IDA PAPER P-2239

THE UTILITY OF SELECTION FOR MILITARY AND CIVILIAN JOBS

Joseph Zeidner
Cecil D. Johnson

July 1989

Prepared for
Office of the Under Secretary of Defense for Acquisition
(Research and Advanced Technology)

INSTITUTE FOR DEFENSE ANALYSES
1801 N. Beauregard Street, Alexandria, Virginia 22311-1772

IDA Log No. HQ 89-34657

DEFINITIONS

IDA publishes the following documents to report the results of its work.

Reports

Reports are the most authoritative and most carefully considered products IDA publishes. They normally embody results of major projects which (a) have a direct bearing on decisions affecting major programs, or (b) address issues of significant concern to the Executive Branch, the Congress and/or the public, or (c) address issues that have significant economic implications. IDA Reports are reviewed by outside panels of experts to ensure their high quality and relevance to the problems studied, and they are released by the President of IDA.

Group Reports

Group Reports record the findings and results of IDA established working groups and panels composed of senior individuals addressing major issues which otherwise would be the subject of an IDA Report. IDA Group Reports are reviewed by the senior individuals responsible for the project and others as selected by IDA to ensure their high quality and relevance to the problems studied, and are released by the President of IDA.

Papers

Papers, also authoritative and carefully considered products of IDA, address studies that are narrower in scope than those covered in Reports. IDA Papers are reviewed to ensure that they meet the high standards expected of refereed papers in professional journals or formal Agency reports.

Documents

IDA Documents are used for the convenience of the sponsors or the analysts (a) to record substantive work done in quick reaction studies, (b) to record the proceedings of conferences and meetings, (c) to make available preliminary and tentative results of analyses, (d) to record data developed in the course of an investigation, or (e) to forward information that is essentially unanalyzed and unevaluated. The review of IDA Documents is suited to their content and intended use.

The work reported in this document was conducted under contract MDA 903 84 C 0031 for the Department of Defense. The publication of this IDA Paper does not indicate endorsement by the Department of Defense, nor should the contents be construed as reflecting the official position of that Agency.

This Paper has been reviewed by IDA to assure that it meets the high standards of thoroughness, objectivity, and appropriate analytical methodology and that the results, conclusions and recommendations are properly supported by the material presented.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	July 1989	Final--June 1988 to July 1989

4. TITLE AND SUBTITLE
The Utility of Selection for Military and Civilian Jobs

5. FUNDING NUMBERS
C - MDA 903 84 C 0031
T - T-D2-435

6. AUTHOR(S)
Joseph Zeidner, Cecil D. Johnson

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Institute for Defense Analyses
1801 N. Beauregard St.
Alexandria, VA 22311

8. PERFORMING ORGANIZATION REPORT NUMBER
IDA Paper P-2239

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
OUSD(A)/R&AT
The Pentagon, Room 3D129
Washington, DC 20301

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT ***(Maximum 200 words)***

The major purpose of this report is to provide military policymakers with procedures for developing and evaluating realistic estimates of costs and benefits of alternative manpower selection and classification policies. Such estimates are needed to make rational choices in allocating scarce resources among strategies for improving organizational productivity.

This report traces the technical development of current decision theoretic selection utility models. The description of selection is extended to introduce more complex classification decision situations that match individuals and jobs to maximize aggregate performance. An overview of the current military system for selecting and classifying manpower is presented along with a discussion of how exclusive focus on predicting validity reduces the efficiency of the ASVAB as a classification tool.

14. SUBJECT TERMS
military personnel selection and classification, aptitude testing, job performance, manpower, classification, selection utility, utility analysis

15. NUMBER OF PAGES
233

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICAON OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	SAR

IDA PAPER P-2239

THE UTILITY OF SELECTION FOR MILITARY AND CIVILIAN JOBS

Joseph Zeidner
Cecil D. Johnson

July 1989

IDA

INSTITUTE FOR DEFENSE ANALYSES

Contract MDA 903 84 C 0031
Task T-D2-435

ACKNOWLEDGMENTS

We wish to express our sincere appreciation for support of this study by making the Army 1984 recruit accession base available and by arranging for the simulations required for the utility analysis to:

Edgar Johnson, Technical Director, The Army Research Institute (ARI),

N. Kent Eaton, Director, Manpower and Personnel Research Laboratory, ARI,

Curtis Gilroy, Chief, Manpower and Personnel Policy Research Group, ARI.

The efforts of Arthur Drucker, formerly at the Army Research Institute, in editing the manuscript are gratefully acknowledged.

We are grateful for the reviews and useful technical comments made by:

Hubert Brogden, Professor Emeritus, Purdue University

Wayne Cascio, University of Colorado at Denver

Curtis Gilroy, The Army Research Institute

Edgar Johnson, The Army Research Institute

Paul Horst, Professor Emeritus, University of Washington

Robert Sadacca, HumRRO

William Sands, Navy Personnel Research and Development Center

Frank Schmidt, The University of Iowa

Richard Sorenson, Navy Personnel Research and Development Center.

This study was performed for the Office of the Under Secretary of Defense (Acquisition)/(Research and Advanced Technology). Technical cognizance for the work, under Task Order T-D2-435, was assigned to the Assistant for Training and Personnel Technology. We are indebted to Earl Alluisi for his contributions to and interest in the effort, and to Jesse Orlansky, Institute for Defense Analyses, for suggesting that the utility of selection and classification procedures be evaluated, for providing valuable information, and for his many useful suggestions on preparing this report.

ABSTRACT

The major purpose of this report is to provide military policymakers with procedures for developing and evaluating realistic estimates of costs and benefits of alternative manpower selection and classification policies. Such estimates are needed to make rational choices in allocating scarce resources among strategies for improving organizational productivity.

This report traces the technical development of current decision theoretic selection utility models. The description of selection is extended to introduce more complex classification decision situations that match individuals and jobs to maximize aggregate performance. An overview of the current military system for selecting and classifying manpower is presented along with a discussion of how exclusive focus on predicting validity reduces the efficiency of the ASVAB as a classification tool.

SELECTION UTILITY FOR MILITARY AND CIVILIAN JOBS

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

AA	Aptitude Area
ACB	Army Classification Battery
AFQT	Armed Forces Qualification Test
AGCT	Army General Classification Test
AI	Aptitude Index
AR	Arithmetic Reasoning
AS	Auto Shop Information
ASVAB	Armed Services Vocational Aptitude Battery
AVF	All-Volunteer Force
CL	Clerical/Administrative
CO	Combat
CREPID	Cascio-Ramos Estimate of Performance In Dollars
CS	Coding Speed
DEP	Delayed Entry Program
DI	Disposition Index
EI	Electronic Information
EL	Electronic Repair
EPAS	Enlisted Personnel Allocation System
F	Finger Dexterity
FA	Field Artillery
FLS	Full Least Squares
g	General Component
G	General Intelligence
GATB	General Aptitude Test Battery
GM	General Maintenance

GS	General Science
GVN	Cognitive Ability
Ha	General Index
Hd	Differential Index
HRA	Human Resource Accounting
LSE	Least Squares Estimate
LP	Linear Program
MAXAACON	Maximize Aptitude Area Score, Constrained
MAXAAFREE	Mazimize Aptitude Area Score, Unconstrained (or Full)
Max-PSE	Maximize Personnel Selection Efficiency
MC	Mechanical Comprehension
MDS	Multidimensional Screening
MK	Mathematical Knowledge
MM	Mechanical Maintenance
MOS	Military Occupational Speciality
MOSLS	Military Occupational Speciality Level System
MPP	Mean Predicted Performance
N	Numerical Aptitude
NO	Numerical Operations
OAE	Operational Allocation Efficiency
OF	Operators/Food
OPM	Office of Personnel Management
OPTAACL	Optimization on Aptitude Area Score, Classification Only
OPTAASC	Optimization on Aptitude Area Score, Selection and Classification
OPTFLS	Optimal Assignment, Full Least Squares Prediction
OPTPRFCL	Optimization on Single Composite Predicted Performance, Classification Only
OPTPRFSC	Optimization on Single Composite Predicted Performance, Selection and Classification

P	Form Perception
PAE	Potential Allocation Efficiency
PACE	Professional and Administrative Career Examination
PAT	Programmer Aptitude Test
PC	Paragraph Comprehension
PCE	Potential Classification Efficiency
PDI	Point Distance Index
PSE	Potential Selection Efficiency
PUE	Personnel Utilization Efficiency
Q	Clerical Perception
RDO	Radial Drill Operator
S	Spatial Aptitude
SC	Surveillance/Communications
SDy	Standard Deviation of Performance in Dollar Terms
SQT	Skill Qualification Test
SR	Selection Ratio
SRQ	Perceptual Ability
ST	Skilled Technical
TC	Tank Commander
u	Unique Component
USAREC	U.S. Army Recruiting Command
V	Verbal Aptitude
VE	General Verbal Ability
WK	Word Knowledge

SUMMARY

The major thesis of this report is that testing saves money because employees selected by valid tests produce more than those selected by other means. Use of Brogden's historic equation developed in 1949 for estimating costs and benefits of a selection program is a means of demonstrating that testing can save money. Similarly, the use of classification tests to assign individuals to specific jobs from a number of available opportunities within an organization also may be demonstrated to save money.

The major purpose of this report is to provide military policymakers with procedures for developing and evaluating realistic estimates of costs and benefits of alternative selection and classification policies using ASVAB. Such estimates are necessary to make rational choices in allocating scarce resources among alternative job entry standards and assignment procedures for improving productivity.

In an earlier report (Zeidner, 1987), major validation studies and meta-analytic summaries were reviewed to assess the effectiveness of selection and classification procedures for predicting job performance in military and civilian settings. The present report traces the technical development of current decision theoretic selection utility models and provides findings of major studies in the literature of utility gains in dollar-valued terms resulting from the prediction of job performance in industry and in civil service. This report also introduces the concept of classification efficiency and cautions that almost exclusive consideration of predictive validity can eventually destroy the usefulness of the ASVAB as a classification tool.

A subsequent report (Johnson and Zeidner, 1989) details classification effects as a component of utility. It provides a taxonomy of the applicant/employee utilization process and methodologies for measuring and improving classification effectiveness.

A final report (Zeidner and Johnson, 1989) with contributed chapters by Roy Nord and Edward Schmitz provides the results of an empirical analysis of productivity gains attributable to simultaneous changes in employee job entry standards (cutting scores) on the ASVAB and in assignment procedures for each of the Army's nine job families. The

implications of the utility findings for future research and military manpower policy are detailed.

The first three reports (on validity, selection utility, and classification efficiency, respectively) are necessary theoretical background for the technically oriented reader to understand the methodologies, assumptions, and estimates made in the fourth report describing the utility analysis. Hopefully, our comprehensive treatment of selection and classification will more readily permit researchers and technical advisers to military policymakers to make more informed judgments in personnel utilization.

A. SELECTION UTILITY

Since the Army's success with testing during World War I, employers were eager to capitalize on the use of standardized tests in the hiring process. Tests were shown to be valid predictors of job performance and perceived as being fair. But to assess the practical impact of findings, practitioners resorted to the use of difficult-to-understand concepts. Starting with the development of the first decision theoretic selection models, a new language began to emerge. In the last decade, thanks to clear empirical findings, it became possible to make economically meaningful "bottom-line" statements on personnel intervention programs designed to improve job performance. In recent years, decision models have become more realistic, comprehensive, integrative and accurate; they permit comparisons to be made among alternative investment strategies on the same basis as other organizational decisions.

The Taylor-Russell (1939) model was the first to show that the context of a selection decision must be considered to evaluate that decision properly. It reformulated the concept of validity away from individual predictions to institutional decisionmaking and redefined the concept of measurement accuracy of predicting decision outcomes.

Brogden's (1949) basic utility formulation, the model that is the basis of all later elaborations, was the first to consider payoff in dollar terms, costs and other external parameters of the selection situation. By the 1960s decision theory was firmly established as the appropriate framework for developing and applying tests.

In the early 1980s a number of practical procedures were developed for obtaining rational estimates of the standard deviation of performance in dollar valued terms, SD_y, the parameter required in the basic utility model, and once believed obtainable only through complex and time consuming cost accounting procedures. Using these practical

procedures, a series of realistic utility studies appeared in the literature. They demonstrated that productivity gains attributable to selection were very large--in the millions of dollars per year.

In the mid 1980s, enhanced basic utility formulations were developed that incorporated financial/economic factors and external employee movement into and out of the workforce. Break-even analysis, a means of simplifying decisions, and risk analysis were introduced to further enhance utility models. A generalized version of the basic utility model was employed to evaluate the dollar value of training. The general version is applicable to all personnel programs intended to improve job performance.

In the last several years a number of studies sought to increase the reliability, understandability and credibility of the SD_y estimating technique. Behavioral based estimates were found to be accurate when validated against external criteria, and credible to managers.

Cumulative research findings strongly suggest that utility analysis will improve organizational decisionmaking. But few utility analysis applications have been used in real-world situations as part of the actual decision process, despite the availability of realistic, comprehensive models. What appears to be much needed now are data on "real" applications of utility analysis intended for use in the process of organizational decision-making.

B. CLASSIFICATION EFFICIENCY

The purpose of personnel classification is to match individuals and jobs in a manner that maximizes aggregate performance. Such classification decisions are a major concern in the military services and of increasing interest in industry and in student counseling. We will refer to the implementation of classification decisions as the assignment process; our generic term for the matching of individuals to either jobs (i.e., military occupational specialty) or to a level within a job (placement) is assignment. In our taxonomy of personnel utilization processes "assignment" is subdivided into "classification" and "placement," and classification" is further subdivided into "hierarchical classification" and "allocation."

Throughout this manuscript we will not distinguish between the decision to attach a military occupational specialty (MOS) to a new soldier, the process which is commonly referred to in the services as classification, and the assignment of an individual to a job.

Thus, we will not distinguish between the assignment of an individual to a MOS and the initial assignment of that individual to a position calling for that MOS at a specific geographic location. The reassignment of a soldier to a new geographic location, at the conclusion of his first tour of duty (a process which does not usually involve a change in MOS), lies outside the scope of our topic.

Traditionally, in selection and placement, only a single job is involved, and can be accomplished with one or more predictors. The outcome is determined by an individual's position along a single predicted performance continuum. Classification decisions provide the basis for assigning a selected pool of individuals to more than one job. As in selection, these assignments can be made on the basis of a single continuum of predicted performance adjusted to reflect job validities or values. When the predictors and the criteria are defined in such a manner that the mean predictor scores and the mean criterion scores have the same rank order across jobs, a hierarchical layering effect that makes a positive contribution to the benefits obtainable from classification will exist. A hierarchical layering effect due to either a variation across jobs of the validities of job specific test composites, or to the value assigned to each job and reflected in predictor score means and/or variances, assures that the assignment process is, at least in part, influenced by hierarchical classification. Classification that does not capitalize on hierarchical layering effects will be referred to as allocation. While hierarchical classification can be unidimensional (e.g., based entirely on a single predictor), allocation requires multiple predictors measuring more than one dimension in the jont predictor-criterion space. Validity is determined individually against each job's performance criterion; the set of job criteria should also be multidimensional. Thus a classification battery requires a separate assignment variable (criterion specific composite) for each criterion, if allocation efficiency is to be maximized. The particular combination of predictors employed out of the total battery plus the specific weight given each predictor varies with each job criterion. In practice, a smaller number of tests than are in the total battery are often used rather than the LSEs (least square estimates) from the total battery, the complete regression equation for all predictors. In the Army, for example, a different unit-weighted, three-test combination or aptitude area composite currently is used in assigning individuals to jobs in each of nine job families.

It is often assumed that the utility of the classification process is a direct function of differential validity. More precisely, differential validity is the level of prediction, using full battery LSEs, of differences among criterion scores. We also use the term in reference to the validity vector for a job having high differential validity, i.e., being more valid for

its own job family than for any other job family. Unfortunately, a simulation study is required to translate the effect of differential validity into mean predicted performance (MPP), that in turn can be readily translated into utility. The utility of a classification battery can be characterized as being directly proportional to the average predicted performance of incumbents in a number of different jobs after an optimal assignment process has been used with quotas taken into account.

When the test content of the selection/classification battery has been fully determined and only the selection of the test composites and weights for use in the selection and/or classification of applicants for each job remains to be determined, the least squares regression weights applied to all tests forming each test composite, the LSEs, provide maximum utility when used in either or both selection and classification. Such composites will not only provide the means of maximizing the average validities across jobs but will also maximize potential allocation efficiency (PAE). The validities of these composites are, of course, the multiple correlation coefficients between the composites and each job criterion measure. No set of composites selected to lower intercorrelations among composites or to increase the variations of composite validities across jobs (as one might mistakenly attempt to do in order to increase PAE) can increase the utility function value as well as the full regression equations based on the total battery. If composites use a reduced number of tests or otherwise are not LSEs, or if jobs are clustered rather than matching each job with their own LSEs, the best composites for selection are not necessarily the best for classification.

Thus the value of using several aptitude areas, rather than one composite, depends upon the presence of potential allocation efficiency (PAE) in the battery from which the tests comprising the aptitude areas were drawn. With the presence of PAE, classification effectiveness would be based on demonstration of specific abilities necessary for different jobs; the set of criterion variables must also be multidimensional. Total human resources would then be more efficiently utilized by capitalizing on scores that indicated differences in the levels of abilities and differences among abilities within each individual (inter- and intra-individual differences).

A serious shortcoming of the current ASVAB composites is their limited ability to differentiate among job families. The same aptitude area used to select individuals specific to an MOS within a job family does nearly as well for MOS in other job families. More specifically, of the nine aptitude areas, only two are more valid for their relevant families than the average validity across families. Thus, while the operational composites are highly

valid, the battery's composites appear lacking in differential validity and one would expect to find little PAE in the ASVAB. With little PAE implied by the lack of differential validity in the composites (an approximate measure of classification efficiency), the benefits obtainable from using more than one occupational composite appear questionable.

There is a choice in the selection of tests for inclusion in the ASVAB: either the focus can be on improving a single general cognitive ability composite or on increasing PAE. For a moderately large operational battery, neither needs to be improved at the expense of the other in the selection of tests. The addition of tests with high PAE does not need to detract from the validity generalization of a number of tests selected to maximize the validity of general mental ability, nor does the addition of tests with high validity generalization capability need to detract from PAE provided by tests specifically selected to maximize PAE. Theoretically, only two to four tests should cover the domains useful for traditional selection, leaving the option of selecting other tests in the battery to further the multidimensionality of the joint predictor/criterion space. Thus, it appears that the implementation of a selection/classification strategy that calls for selecting some tests to maximize the magnitude of validity coefficients and other tests to maximize PAE can achieve most of the PAE possible while losing little, if any, capability for validity generalization. Once a battery is selected, the same weights are best for achieving either the maximum average validity in accordance with validity generalization or to maximize PAE. Of course, the maximum PAE would not be achieved unless a maximum allocation process (with respect to both variables and procedures) is used.

The possibility of fully benefiting from a deliberate consideration of PAE, with little or no decrease in average validity as a consequence, depends upon the following conditions: (1) (most important) whether the battery and composites are fixed (already determined); (2) whether the selection/classification process is accomplished in one or two stages (simultaneously or sequentially); (3) which optimal selection/classification procedure is being utilized to implement assignment to jobs (an LP type program); and (4) whether job families are appropriately structured (smaller differences among LSEs for jobs within families and larger differences between LSEs for jobs across families, or, ideally, one LSE for each job).

Selection and classification are both essential parts of a personnel utilization process. Both must be considered in estimating the utility obtainable from the process. The use of mean predicted performance (MPP) provides the common thread that links selection and classification to utility. Estimates of utility resulting from personnel utilization

involving both selection and classification procedures must be based on a specified assignment process that considers the effects of both selection and classification.

The potential benefits of optimal assignment are usually not realized because of the nature of the operational assignment process used in practice. The traditional assignment approach used in the military, for example, is a two stage process: selection is first accomplished based on AFQT entry level recruitment standards; then classification is accomplished on the selected group through the use of aptitude area composites. Benefits, however, are maximized through the use of a single stage selection/assignment process (i.e., multidimensional screening, the *MDS* algorithm that integrates the effects of both selection and classification). Using the MDS model, both processes are accomplished simultaneously through the use of different cut scores optimized for each job family predictor composite. An optimal selection/classification process most probably has never been used in any operational context.

We describe means of correctly identifying a personal utilization process as either selection, classification, placement, classification/allocation, hierarchical classification or some combination of these procedures. We emphasize both the importance of the selection/assignment process in obtaining maximum benefits from a battery and in estimating personnel utilization benefits as the first step in estimating utility.

We also define and describe means of defining and measuring potential allocation efficiency (PAE), potential classification efficiency (PCE) and potential utilization efficiency (PUE). The total selection, classification and placement process, individually or in combination, is termed the "personnel utilization decision process." Classification efficiency may be subdivided into two effects: allocation efficiency and hierarchial classification efficiency. All classification efficiency not due to hierarchical layering effects, when heterogenous validities and/or values are assigned to jobs and also reflected in the predictor variables used in the assignment process, is attributable to allocation efficiency. When the classification test battery is unidimensional, no allocation benefit can exist; the assignment process consists entirely of hierarchical classification. If all assignment variables (e.g., aptitude areas composites) have equal means and variances, the classification process is pure allocation since no means for a hierarchical classification process to capitalize on hierarchical layering is present. However, when hierarchical layering of validities or job values exist and are reflected in the predictors, and the joint predictor-criterion space is multidimensional, the classification process includes both hierarchical classification and allocation processes. When both hierarchical classification

and allocation are present in the same process, their effects are so confounded as to make them difficult, if not impossible, to separate.

OVERVIEW--SELECTION UTILITY

Improving U.S. business productivity in providing goods and services is a subject of major concern shared by policymakers in government, industry and labor. The causes attributed to declining productivity in the last decade are numerous. But no matter what the causes, both management and labor are now actively working for ways of improving productivity. Nearly everyone acknowledges that people are the key to productivity, and that productivity gains depend greatly on matching the attributes of people with the demands of the job.

From the time of the Army's success with the Alpha test during World War I, employers were eager to capitalize on employment testing in the hiring process. Tests were shown to be valid predictors of job performance and perceived as fair. However, since the passage of Title VII of the Civil Rights Act of 1964, employers have become extremely cautious in the use of tests because they have feared charges of discrimination, unfairness, and adverse impact.

Return to full employment testing may depend, in part, on societal acceptance of research findings on test "fairness" showing tests predict job performance with equal accuracy for both minority and majority groups. Return to full testing may depend even more strongly on societal awareness of the large nationwide economic benefits to be realized through employment testing.

But to assess and communicate the practical impact of testing, experts have resorted to the use of difficult-to-understand statistical concepts or behavior terms. Starting with the development of the first decision theoretic selection models, a new language began to emerge. In the last decade, based on clear empirical findings, it became possible to make economically meaningful "bottom-line" statements on personnel intervention programs designed to improve job performance. Today productivity gains attributable to selection can be expressed in dollar-valued terms comparable to other financial investments made by organizations.

The major purpose of this report is to trace the technical development of current decision theoretic selection utility models. Current utility models represent a shift away

from classical or traditional models of individual prediction accuracy. They provide the means of determining institutional productivity gains, in dollars, which would result from using a predictor in a specific decision context.

A case in point is the widespread interest in cost-effective military manpower utilization policies for establishing test entry standards for enlistment and for making job assignments (classification). Interest extends beyond military policymaking as a public policy issue because the standards used have social, political, economic and national security implications. In the last thirty years, for example, changing military manpower policies concerning the "quality" of enlisted personnel have resulted in three distinctive capability levels of armed forces, ranging from high to low.

The military services cannot, of course, fill all of their manpower requirements with high-quality recruits because it would be prohibitively expensive (even turning to the draft). The services determine by formal standards [e.g., scores on the Armed Forces Qualification Test (AFQT), education, criminal record] and informal standards (e.g., incentives given to recruiters) who will be permitted to enlist. These formal and informal standards, along with the resources devoted to recruiting and the values and attitudes of the youth population, determine the quality mix of the services as a whole and for specific job specialties.

Although attracting more high-quality recruits requires expending more resources, costs must be balanced against the level of performance of high-quality recruits. This economic concept of cost/performance tradeoffs is embedded in the core of selection and classification utility models. Exactly the same cost/performance tradeoffs pertain in hiring employees in industry as in government.

Testing can save money because employees selected by valid tests produce more than those selected by other methods. Use of Brogden's historic equation developed in 1949 for estimating costs and benefits of a selection program has been a means of demonstrating that testing can save money. How much is saved depends on the predictive efficiency of the selection device, the selection ratio (the proportion of applicants hired), and two recently applied situational variables of importance--the variance of dollar-valued performance to the organization and costs associated with testing.

In Chapter 1, we begin with a description of the All Volunteer Force (AVF) and the use of the AFQT to determine entry into military service. The quality of the force today stands in stark contrast to the force in earlier years of the AVF. The Army's Chief of Staff

told Congress in 1980 that he led a "hollow Army". Such military force quality changes point to the importance of developing and evaluating alternative manpower utilization policies on a systematic basis.

Schmidt and Hunter (1981) noted that the once prevalent "theory of low utility" could be used by test "fairness" advocates to support the view that selection procedures had little impact on productivity and thus lend support to the view that test selection procedures could be ignored to achieve other goals, including a racially representative work force. However, on the basis of utility findings, Schmidt and Hunter (1981) concluded that one possible reason for the recent decline in the rate of growth in productivity can be traced to the abandonment of tests by organizations in response to pressure from the federal government.

Utility analysis became theoretically possible with Brogden's (1949) historic equation for estimating costs and benefits. Brogden and Taylor's (1950) classic article argued further for supporting the use of a dollar criterion to link validity with a firm's objective of making money. But Brogden's equation required a scale translating selection validity into dollar-valued performance. Thirty years later, a practical means was developed for estimating the standard deviation of employee job performance in dollars that could replace cost accounting methods. Utility analyses then began to appear in the literature. Cascio's (1987a) book on costing human resources strongly endorses the view that the language of business is dollars, not correlation coefficients. Decisionmakers take human resource costing in dollars quite seriously.

Cronbach and Gleser (1965), in their influential book, firmly established decision theory as the appropriate framework for developing and applying tests. They demonstrated that every decision problem must be specified and these specifications must be used to determine the appropriate mathematical model. Decision theory, they stated, is distinguished from simpler models by the fact that it is built of concepts that are often neglected--the set of alternatives, the costs, and the possible outcomes. In working out formal solutions, however, it becomes necessary to neglect certain key concepts and to introduce strong assumptions. Cronbach and Gleser say that traditional theory views tests as measuring instruments intended to assign accurate minimal values to some quantitative attribute of the individual. It stresses precision of measurement and estimation. In decision theory, or using decision theoretic models, however, a quantitative estimate is not the real desideratum; rather a choice must be made between alternatives. The appropriateness of the

traditional model's interpretations of validity was challenged earlier by Taylor and Russell (1939) and Brogden (1946a, 1949).

The principal concern of utility analysis, then, is the determination of value or payoff resulting from a program's consequences. Two different conceptual approaches, traditional and decision theoretic, have been taken to measure payoff. Chapter 2 provides descriptions of selection utility models and their associated payoff measures in the historical literature.

A number of psychometric interpretations based on simple computational operations performed on the validity coefficients have been suggested by traditional measurement proponents, including Kelley's (1923) coefficient of alienation, Hull's (1928) index of forecasting efficiency and, in the 1930s and 1940s, the popularly accepted coefficient of determination (r^2_{xy}).

If such traditional interpretations of the "low efficiency of validity coefficients" were to be accepted as an appropriate utility index, personnel testing would indeed be in trouble. But these traditional approaches, evaluated from a decisionmaking perspective, have no direct bearing on the value of a selection program because they do not properly consider the external context of the selection decision nor do they make appropriate assumptions concerning the utility function. The traditional validity approach is confined to maximizing correct hires and rejections while minimizing erroneous hires and rejections (i.e., combining measures of squared deviations from a predicted linear function). Thus, the traditional approach implicitly treats all possible selection-decision activities as equally good or bad.

Brogden (1946a) provided the most widely accepted current decision theoretic interpretation of the validity coefficient. He showed that the validity coefficient is a direct index of selective efficiency. The validity coefficient could be expressed as the ratio of the mean performance of those selected, using the predictor, to the mean performance of those selected if selection were done on the basis of the criterion (e.g., a predictor with a validity of 0.50 would produce 50% of the gain resulting from a perfect selection device).

The Taylor-Russell decision theoretic approach measures utility as a function of the success ratio by reformulating measurement accuracy to that of accuracy of predicting decision outcome. The organizational goal becomes that of identifying the largest number of potentially "successful" and "unsuccessful" employees in the applicant group. The approach goes beyond the validity coefficient to consider two additional parameters--the

selection ratio and the base rate (the percentage of applicants who would be "successful" without the use of the selection procedure). In including the context of the situation, this model was the first to show the complex interactions among the validity coefficient, the selection rate and the base rate in affecting decision outcomes. It uses a dichotomous criterion, however, that ignores real differences in performance. The decision as to where to draw the dividing line between successful and unsuccessful employees is often arbitrary.

In Brogden's (1946a) formulations, he also showed that utility can be expressed as a function of increase in a continuous criterion score. For any given arbitrary specified cutoff on a predictor, the higher the validity, the greater the increase in the mean criterion score for the selected group over the score of the total group. This is a practical utility index that is more generally applicable than the Taylor-Russell model. Because Brogden's index is expressed in standard score units rather than in dollar terms or production quantities, it may be difficult for decisionmakers to use the index on the same basis as other economic measures in making investment decisions. However, Brogden (1949) expanded the components of his 1946 equation to consider payoff explicitly in terms of dollars, costs and other external parameters of the selection situation.

The derivation of Brogden's (1949) utility equation is here reviewed, since his equation is the basis of all future utility model extensions. It is especially noteworthy for this overview:

$$U = TNr_{xy}SD_y Z_{xs} - NC$$

where U = the total utility gain of a selection program; T = the duration, in number of years of a selection program's effect on performance; N = the number of those selected in the applicant population; r_{xy} = the validity of the predictor; SD_y = the standard deviation of job performance in dollars; Z_{xs} = the average predictor score of those selected in the applicant population standard score; and C = the cost of selection per individual.

Brogden's utility model requires only the commonly accepted assumption of linearity between predictor and job performance. Since the model incorporates all essential components of utility (quantity, quality, and costs) and provides an incremental index that permits comparisons to be made between existing and new selection programs in dollar terms, it is superior to all previous decision theoretic selection utility models.

Hunter and Schmidt (1982) pointed out that despite the availability of Brogden's equation since 1949, utility analysis did not receive widespread attention until recent years. They attributed this lack to three factors: concerns about statistical assumptions exactly

fitting the linear model; the erroneous acceptance of the viewpoint that validity is situationally specific, thus requiring expensive validations for each new application; and the belief that difficult cost accounting procedures were necessary in estimating the dollar value of employee performance. Schmidt and Hunter (1982) provided analyses that helped ameliorate concerns about utility studies and also began to demonstrate through practical empirical investigations, large work force productivity gains attributable to valid selection procedures.

Chapter 3 details methods of measuring the payoff scale in dollar terms, SD_y, the parameter most difficult to obtain in practice. During the last decade, new procedures were developed that require only the judgment of experts in estimating values for computing SD_y. While there is only one basic utility model, these new behaviorally based approaches offer alternative procedures for measuring the payoff scale.

Brogden and Taylor (1950) suggested the use of cost accounting procedures to develop a dollar criterion that measures the contribution of the individual to overall efficiency of the organization. Roche (1961) conducted a field study that directly applied Brogden and Taylor's cost accounting elements to beginning level radial drill operators based on "standard costing" procedures. Roche concluded that while the study clearly demonstrated that a dollar criterion could be developed, many estimates and arbitrary allocations enter into cost accounting. Others have criticized the cost accounting approach in terms of its complexity, its procedural difficulties and the effort entailed. Cascio (1987a) also concluded that cost accounting systems focus on the costs and benefits of units of products, not units of performance; many estimates are needed, and the objective of cost accounting data may become suspect. Nevertheless, Greer and Cascio (1987) stated that a cost accounting approach to the estimation of SD_y "remains as the conceptual standard of comparison."

The difficulties in applying cost accounting eventually led to the development of an entirely different and greatly simplified approach to estimating SD_y. This new procedure, the global method of obtaining rational estimates, was developed by Schmidt, Hunter, McKenzie and Muldrow (1979) and awakened a renewed interest in utility analysis. The procedure is global in the sense that one obtains overall dollar estimates from supervisors of the value of goods and services for all employees performing at the 15th, 50th and 85th percentiles, in contrast to identifying and weighing separate task components underlying total dollar-valued performance for each employee separately. If job performance in dollar terms is normally distributed, then the difference between the value to the organization of

the products and services produced by average employees at the 50th percentile and those produced by employees at the 15th percentile (or 85th) in performance is equal to SD_y.

Most of the research accomplished to date has been directed at evaluating the reliability and validity of global estimates since it is among the first behaviorally based methods to be developed, is most frequently used in utility analysis, and generates very large SD_y estimates. Arguments critical of this method principally concern its accuracy, its variability, the normality of the underlying distribution, and the judgment process involved in determining payoff values.

An alternate approach, the Cascio-Ramos estimate of performance in dollars (CREPID), returns to the older tradition in psychology of measuring job performance directly using carefully constructed behavioral rating scales based on job analysis results (Cascio & Ramos, 1986). The method assumes that if an organization's compensation program reflects current market rates for jobs, then the economic value of each employee is reflected best in his or her salary. The method also assumes that all significant aspects of performance can be detailed in the rating scales for principal activities (a percentage of salary being assigned to each activity), and that the rating scales can properly reflect individual differences in performances. Since the economic value of goods and services is about twice the average salary, the CREPID approach, not surprisingly, yields much lower estimates--about half the size of the global method. The process as a whole is systematic, explicit, and understandable and consequently may be credible and acceptable to users.

Three other methods also are described in this chapter. Schmidt and Hunter (1983) developed an estimating procedure based on proportional rules (i.e., SD_y equals 40% of average salary or 20% of mean output as lower bound estimates). Eaton, Wing and Mitchell (1985) developed two techniques, "superior equivalents" and "systems effectiveness" for estimating dollar value of performance that appeared to be useful in certain job contexts, e.g., where performance is largely a function of a team leader or where employees operate very complex, expensive equipment focal to the productivity of a costly system.

A number of empirical comparisons of alternative SD_y estimates are described. These studies make comparisons of inter-rater variability, consistency with other measures, normality of the underlying performance distribution, nature of the dimension being measured, accuracy of estimates and acceptability by decisionmakers. While most studies show inconsistencies among methods, overall results offer encouragement because some

estimating procedures show good accuracy when compared to the accuracy of external validation procedures, the degree of multimethod convergence of estimate values, and the face validity or credibility of some procedures.

Chapter 4 reviews a number of empirical studies that demonstrate the savings that can be expected from improved selection procedures in realistic decision contexts. Seven utility analyses were selected as being significant or exceptional in some ways (e.g., determining utilities for most white-collar jobs in the federal government, estimating nationwide productivity gains resulting from allocation procedures, and employing risk analysis for the first time in utility analysis).

Results of all studies showed very large utilities--gains in terms of the millions of dollars that can accrue to organizations annually from valid selection procedures. An example is the frequently cited study of selection utility for computer programmers (Schmidt et al., 1979). The Programmer Aptitude Test (PAT) was selected as the predictor for the computer programmer job because previous meta-analytic research showed the very high validity of 0.76 and also found that validity was essentially constant across different organizations. Therefore, it becomes possible to estimate PAT utilities in both the federal government and the economy as a whole, given an assumed testing cost for PAT of $10 per examinee.

Building all realistic factors into Brogden's utility formulation, at one extreme, if SR is 0.80 and the procedure PAT replaces has a validity of 0.50, the productivity gain is $5.6 million for one year's use of the test in hiring 618 new programmers in the federal government. At the other extreme, when SR is 0.05 and the previous procedure has zero validity, the one year's productivity gain is $97.2 million.

If the entire incumbent population of 18,498 programmers in the federal government at that time had been selected by PAT with a validity of 0.76 in place of a procedure with a true validity of 0.30 and a SR of 0.20, then the productivity gain for one year's use would have been about $1.2 billion; expanding this example to the economy as a whole, the productivity gain would have been $10.78 billion, assuming that the number of job seeking programmers far exceeded the number of jobs. The productivity gains for an organization, however, cannot be extrapolated in a simple way to all (programmer) jobs making up the national economy because of a variety of factors, including competition and the nature of the labor market.

In addition to the usual statistical assumptions, the authors of the study note that productivity gains depend on the assumption that selection proceeds from the top-scoring applicant downward until the SR has been reached (i.e., the analysis is based on optimal selection procedures). An additional assumption is that all applicants offered the job accept them since rejecting hiring offers by applicants would have the effect of increasing the SR. Also an implicit assumption is made that the organization's applicant pool is a representative sample of the potential applicant pool. It is apparent, then, that an organization must be in a position to recruit and hire the most qualified applicants to obtain the full economic benefits of a valid selection procedure.

The Schmidt et al. (1979) study is significant because it was the first to demonstrate realistically the application of decision theoretic utility equations and the first to show the magnitude of potential productivity gains attributable to valid selection procedures. Later studies reported in this chapter are increasingly more comprehensive and realistic. For example, they incorporate economic considerations (Burke and Frederick, 1986) and the effects of employee flows and risk analysis (Rich and Boudreau, 1987). The collective results of the studies not only show that very high productivity gains can be attributed to selection, but that utility analysis contributes to better understanding of the decision context and is useful in deciding among competing investments.

In Chapter 5, new uses and extensions of the basic utility model are described. Schmidt, Hunter and Pearlman (1982) generalized Brogden's equation to make it applicable to any type of personnel program designed to improve performance. They showed that the product of r_{xy} and Z_x in the basic equation may be replaced by d_t, the true difference in job performance (correcting for criterion unreliability).

As Landy, Farr and Jacobs (1982) noted, it now appears possible to view the entire system by which organizations select, train, place, and motivate employees from a utility performance perspective because the object of these interventions is to increase the mean performance of the potential workforce.

Mathieu and Leonard (1987) conducted an operational empirical evaluation of a training program in supervisory skills on the performance ratings of bank supervisors. They used a sophisticated expanded version of Schmidt et al.'s (1982) formulation to consider economic factors, employee flows, and diminishing effects of training. The results were compelling not only in terms of dollar savings but also from the standpoint of information provided for managerial decisionmaking.

Boudreau (1984) applied break-even analysis to selection utility, pointing out that instead of estimating the level of expected utility for each alternative, a simpler procedure would be to identify the break-even values critical to making a decision. For example, in evaluating the utility of the Programmer Aptitude Test (PAT), validity and cost emerged as the only two relevant decisionmaking variables in Schmidt et al.'s (1979) study. Because the PAT is more valid *but* more costly than random selection, two decision options pertain: random selection or selection by means of the PAT. Break-even analysis readily provides the break-even value the PAT needs to exceed in order to be the alternative of choice. Since the aim of break-even analysis is to produce only the basic information needed to make the decision, precise SD_y estimates may not be necessary. Because reported productivity gains in other studies had been uniformly high, if only the break-even points in these studies had been computed, it appears nearly certain that decisions concerning whether or not to adopt programs would have been unaffected.

Boudreau (1983a) extends utility formulas by incorporating three financial and/or economic considerations: variable costs, taxes, and discounting. He defines the payoff function as net benefits or the difference between sales value and service costs to make the definition more consistent with other financial investments. The value in incorporating economic considerations in utility models is that they provide a more defensible and realistic utility definition. While Boudreau's formulation can often lead to lower utility estimates, such estimates remain substantial and provide compelling evidence of the value of personnel programs.

On the other hand, failure to consider the effects of employee flows may underestimate utility estimates even more. Most early utility models assumed that a selection program was offered to one group of applicants, and provided the utility of adding the one-treated cohort group, i.e., a single group of selected applicants, to the existing work force. Boudreau (1983b) extends utility models by incorporating the flow of employees into and out of the work force. In 1985 Boudreau and Berger developed a more general external employee model. This later model provides a framework for even further integration and expansion of models to increase realism and accuracy.

Chapter 6 addresses current issues in utility analysis. The strength of utility analysis is the degree of realism it embodies compared to simpler models. But all utility analyses make simplifying assumptions and estimate parameters that cannot be measured precisely. Some assumptions limit utility to a relatively narrow subset of decision

situations, even though they may not be unrealistic. Other assumptions may be difficult to accept or may be irrelevant to decisionmaking.

The determination and interpretations of productivity gains resulting from personnel interventions are, in general, consistent with the economic way of thinking. But there are some important differences. For example, in determining economic marginal utilities of a production function, a broad range of relevant organizational costs enter into consideration. Production theory implies that a factor's marginal product is dependent on the relevant amount of other factors with which it is combined. The same holds for measuring individual productivity. In selection utility, by way of contrast, productivity gains are estimated only for the effects attributable to a selection program, holding all other factors constant. Although job performance measures link individuals with their jobs, actual productivity will be influenced by other factors (e.g., the number and quality of co-workers, the equipment to do the job, etc.). Additionally, individual measures of performance will not provide the data needed to determine the best combination of production factors.

Although present utility models incorporate a number of organizational activities, they still do not include many important interacting internal and external organization phenomena. Utility estimates pertain to potential gains of *future* productivity increments attributable to the intervention, but assume all other *present* organizational parameters will be constant or stable. For example, while labor market conditions significantly impact the pattern of acquisition and retention of employees, characteristics of the future applicant group and costs of employees are considered constant, while only the effects of the program are considered (e.g., the effect of higher-quality employees on productivity).

The concept of differential validity and its contribution to classification efficiency is described. In this context, we find that the current Armed Services Vocational Aptitude Battery (ASVAB) composites have high validity, but show little differential across job families (most composites designed for a specific job family have even higher validities for one or more other job families); the lack of differential validity for most of the composites provides little hope that a more precise evaluation of the composites in the context of a representative set of Army jobs would show an acceptable level of potential allocation efficiency for this particular set of test composites. Thus, the benefits obtainable from using more than one occupational predictor composite are disappointingly low without capitalizing on hierarchical layering effects. Nevertheless, it might be possible to find and

exploit the presence of PAE in future operational batteries designed expressly for that purpose.

The cumulative findings of utility analysis research strongly suggest that it will improve organizational decisions. They include many demonstrations of large potential productivity gains of personnel programs and many illustrations of the uses and advantages of new elaborations of the basic utility model. But what now appears to be much needed are data on "real" or operational applications of utility analysis as an integral part of the process of organizational decisionmaking. Such applications would foster the theory and technology of utility analysis and help institutionalize its use.

CHAPTER 1. THE USE OF TESTING FOR SELECTION DECISIONS

A. INTRODUCTION

This report is the second of two reports evaluating the utility of standardized testing. The first report is concerned with the validity of selection and classification procedures for predicting job performance in military and civilian settings (Zeidner, 1987); the second report (the present one) considers the economic benefits of predicting job performance. (See Appendix A for a glossary of terms.)

The present report reviews the concept and measurement of utility in selection, placement and classification, describes methods of estimating the economic benefits of job performance, and examines the costs and benefits of selection for a variety of jobs.

A major technical focus of this report and of two subsequent reports (Johnson and Zeidner, 1989; Zeidner and Johnson, 1989) concerns the methodologies employed in determining the economic value of increases in soldier performance attributable to the combined effects of alternative classification and assignment policies (utilization decisions) following selection. In classification, utility gains are dependent on differential validity, the level of prediction of differences between job performance measures. Thus the evaluation of classification policies must always be considered together with assignment, i.e., the person-job matching procedure. A simulation study (rather than an analytic study used in selection), employing an allocation procedure, is required to translate the effects of differential validity into utility.

In determining classification utility, we develop the argument that mean predicted performance (MPP) criterion standard scores can be readily transformed into a common underlying metric for expressing value across multiple jobs, as is true for unidimensional selection. If SD_y, the standard deviation in dollar-valued performance, is used for expressing the value of each job, MPP can be transformed into dollar terms for both selection and classification. In our utility analyses we use the 40% of average salary proportional rule to conservatively estimate SD_y. By expressing MPP in dollar terms, we are able to link benefits and costs to measure the utility of classification policies.

While value may also be expressed in terms of relative importance and used as the common metric across jobs for classification purposes, it does not provide a means of linking benefits and costs to measure utility. For example, subjective estimates of the relationship between ability levels and importance levels for each job can be determined and used for determining optimal classification. However, the benefit of such a classification system cannot be expressed relative to the cost of recruiting applicants, e.g., should job standards be raised or lowered considering net productivity?

Interest in cost-effective military manpower utilization policies, including selection and classification, extends beyond military policymakers. Manpower utilization is a concern of public policy because of its social, political, economic, and national security implications. In the last thirty years, changing military manpower policies directly impacting on the "quality" of enlisted personnel have resulted in three quite distinctive types of armed forces, ranging in capability from high to low.

The Armed Forces Qualification Test (AFQT), a composite of verbal and math tests of the ASVAB, is the selection device used to determine entry into military service, and also as a measure of general mental ability. It is used by Congress and the military as an index of manpower "quality". About one half of all applicants rejected for military service are denied because of failure to achieve the required AFQT cutting score.

While the AFQT is the single psychometric measure used by all services for determining acceptance into the military, aptitude composites or combinations of ASVAB tests unique to each Service are used for classifying recruits for various types of technical training and subsequent assignment to jobs. In the Army, for example, the ten tests of ASVAB are combined into nine aptitude area composites, such as clerical or administrative, combat, electronics, and general maintenance. Aptitude area composites are used in matching soldiers to specific Army jobs or military occupational specialities (MOS) from among the 260 or so entry level MOS which are clustered or grouped into MOS job families or career management fields (CMF) comparable to civilian job family taxonomies.

For fiscal year 1987, the All Volunteer Force (AVF) met or exceeded its recruiting goals for the eighth consecutive year without lowering its quality standards. Of the 315,000 recruits entering military service, 93% had a high school diploma and 95% scored in the top half of the AFQT. These results are far above the graduation rate (about 75%) and median AFQT score for the American youth population as a whole--a record not only of high quality of youth attracted to military service, but of high retention as well.

Nevertheless, manpower policymakers are beginning to express considerable doubt that they can sustain this manpower quality level in future years.

There is clear cause for pessimism. Recent restraints imposed on military spending because of the economy and budget deficits have again resulted in military compensation lagging behind civilian compensation. With the decline of the youth population into the mid-1990s, there will be increased competition for high-quality male recruits among the services, educational institutions and the private sector. Military personnel specialists are increasingly worried that the sophisticated technologies embedded in today's weapons systems may already extend beyond the abilities of enlistees that the services could realistically expect to attract and retain. Accordingly, advocates of the draft, including influential senators, agree that manpower costs eventually can be met only through a return to conscription.

The quality of the force today stands in stark contrast to the force in the early years of the AVF. The services were widely perceived as being dispirited, comprised of a large percentage of poor-quality high-school dropouts difficult to train, motivate and discipline. Moreover, pay was poor and the public was largely indifferent to the plight of the services. No wonder, then, that General Edwin C. Meyer, the Army's Chief of Staff, told Congress in 1980 that he led a "hollow army." With regular increases in pay starting in the early 1980s, along with renewed political and public interest, manpower quality began to climb steadily to the current high level.

Prior to the AVF, the draft brought men into the Army during the Vietnam War while the potential of being drafted encouraged some to enlist into the other services. While the quality of manpower tended to reflect the overall youth population, except at higher quality levels due to liberal deferments for educational and employment purposes, reenlistment rates were disappointingly low and the services became dependent on the draft and first-term enlistments to meet manpower demands. Manpower policy influences on armed forces quality and retention have been clear-cut. In the 1960s, policies resulted in armed forces of average quality and low retention; in the 1970s low quality and low retention; and in the 1980s high quality and high retention. These results lend emphasis to the importance of decisionmakers' ability to cope with change effectively by generating and evaluating alternative manpower utilization policies on a systematic basis.

We start with an overview of traditional and decision-theoretic conceptual approaches to utility.

B. PRODUCTIVITY, PERFORMANCE AND TESTING

Improving U.S. business productivity in providing goods and services was a subject of major concern shared by governmental policymakers, industrial and labor leaders, workers and scientists during the last decade. The causes attributed to declining productivity are numerous, complex and interrelated. But no matter what the causes, both labor and management are now actively working for ways of improving productivity.

One such area of universal concern is personnel costs. Managers desire to keep a tight reign on labor costs as a direct and promising means of increasing the profit side of business. Usually, the first defense against "excessive" personnel costs is to impose personnel freezes, reduce work force, and tighten budgets. Managers feel that fat can be cut in the personnel area without much impact on organizational performance, since estimates indicate that workers are only about 50% "productive". But such actions, when taken, invariably reduce organizational performance and result in reduced morale and undesirably high employee turnover.

Everyone acknowledges that people are the key to productivity and that overall efficiency ("doing things right") and effectiveness ("doing the right things") depends greatly on matching attributes of people with demands of jobs. From the time of the Army's success with the Alpha tests during World War I, employers were eager to capitalize on the use of standardized tests in the hiring process as an effective means of increasing work force productivity. Tests were not only shown to be valid predictors of training and job success, but they also were perceived as being both convenient and fair.

For nearly a half-century personnel testing continued to be seen as a vital human resource activity contributing to productivity in industrial, governmental, and military settings. However, since the passage of Title VII of the Civil Rights Act of 1964, employers have become extremely cautious in using tests to make hiring and other personnel decisions. They had feared charges of discrimination, unfairness, and adverse impact. Friedman and Williams (1982) analyzed the use of employment testing over the last two decades and attributed its decline to employers' attempts to reduce their vulnerability to litigation under the provisions of the Uniform Guidelines on Employee Selection Procedures as applied by the Equal Employment Opportunity Commission (EEOC, 1978).

Friedman and Williams believe that we may now be seeing a slow return to testing by employers partly because written tests can be more readily defended than alternative

procedures. Return to full employment testing may also depend on the acceptance by the public and the judiciary of findings from numerous regression model studies on test fairness results that conclusively show tests predicting job performance with equal accuracy for both minority and majority groups (Bartlett, Bobko, Mosier & Hannan, 1978; Boehm, 1977; Hunter & Hunter, 1984; O'Connor, Wexley & Alexander, 1975; Schmidt & Hunter, 1981). Return to full testing may depend even more strongly on societal awareness of the magnitude of economic benefits to be realized through employment testing as shown by a growing number of utility studies.

Schmidt, Hunter, McKenzie and Muldrow (1979) pointed to the emphasis given to the practical utility of selection procedures in case law and the failure of personnel psychologists to realize the importance of valid selection procedures on work force productivity (economic utility). Schmidt and Hunter (1981) argue convincingly against the once prevalent "theory of low utility" which held that employee selection procedures had little impact on the performance and productivity of the resultant work force. The theory thus could be used to support the view that test selection procedures could be ignored or safely manipulated to achieve other goals, including a racially representative work force. Schmidt and Hunter and their colleagues were the first to stress the impact of high test performers on organizational productivity as shown in utility analyses, expressing the benefits of selection in terms of dollar savings. On the basis of utility analysis findings, Schmidt and Hunter (1981) concluded that one possible reason for the recent decline in the rate of growth in U.S. productivity can be traced to the abandonment of tests by organizations in response to pressure from the federal government.

Utility analysis became theoretically possible by Brogden's (1949) basic equation for estimating costs and benefits. He demonstrates that if the criterion is expressed in cost accounting terms the dollar savings of a selection program can be estimated. Brogden writes:

> Testing can save money. Savings result because workers selected by valid tests produce more than workers selected by less efficient methods. How much is saved depends on two factors: (1) the effectiveness of the selection instruments in predicting efficiency on the job and (2) the percentage of applicants who must be chosen. The first of these has received much attention, and much effort has properly gone into the development of tests and interviews which will have the highest possible validity under the given circumstances. The importance of the second factor has not been so universally recognized, although . . . great increases in production can be achieved with a decrease in the selection ratio . . . (p. 171)

Brogden and Taylor (1950) argued for the use of a dollar criterion to serve the functions of choosing the "best" battery from among a number of experimental tests and estimating the validity of the battery. The authors state that the criterion should measure the contribution of the individual to the overall efficiency of the organization. Their reasoning leads to the consideration of the objectives of the organization:

> The general objective of industrial firms is to make money. Monetary saving, being the objective of the organization, is the logical measure of the degree to which on-the-job activity of the individual contributes to or detracts from this overall objective. Only after we have succeeded in evaluating on-the-job performance in these terms can we be sure that our criterion measures conform to the objectives of the organization. It seems apparent that examination of the way in which a given employee affects overall efficiency requires that we determine the way in which his on-the-job activities produce objects or services of monetary value and the ways in which his errors, accidents, spoilage of materials, etc., result in monetary outlay.
>
> It is believed, however, that unless criterion elements are of such a nature that they can be expressed in dollar units, their use as criterion measures cannot be directly justified and do not satisfy the requirement of logical face validity previously discussed. (pp. 139-140)

But the equation called for in Brogden's (1949) classic article required a scale translating selection validity into dollar-valued performance. By the mid-1970s feasible methods were developed for estimating the standard deviation of employee job performance in dollars, the value needed for Brogden's equation, and utility analyses then began to appear in the literature (Hunter & Schmidt, 1982; Schmidt, Hunter, McKenzie & Muldrow, 1979).

Commenting on these developments, Hunter and Schmidt (1983) write:

> Applied psychologists have conducted research on a variety of organizational interventions aimed at increasing employee job performance and productivity (Katzell & Guzzo, 1983). The usefulness of this research for business and government has often been bounded by constraints: (a) the extent to which findings can be made definitive; and (b) the extent to which the impact of findings can be stated in administratively and economically meaningful terms. To render findings definitive, one must reconcile the apparently conflicting results of different studies. To assess the practical impact of findings, one must translate such arcane psychological jargon as "$p < .01$" into economically meaningful statements such as "a 10% increase in output" or "a reduction of $100 million in labor costs." Recent advances have been made in both areas under the rubrics *meta-analysis* and *utility analysis*. (p. 473)

Cascio (1987a) in his book on costing human resources also strongly endorses the approach first promulgated nearly 40 years earlier:

> For some time now, I have had the uneasy feeling that a lot of what we do in the personnel or human resource management field is largely misunderstood and underestimated by the organizations we serve. In part, we in the field are responsible for this state of affairs because much of what we do is evaluated only in statistical or behavioral terms. Like it or not, *the language of business is dollars, not correlation coefficients.* (p. ix)

C. DECISION-THEORETIC UTILITY APPROACH TO SELECTION

Personnel selection is a decisionmaking process--one of choosing among alternative courses of action. At the simplest level, it is the decision to hire from a large number of applicants those applicants who are most likely to perform well on the job. The decision process may be as simple as rank-ordering applicants on the basis of desired attributes on a battery of standardized tests and either selecting the highest scoring individuals or rejecting those below a minimum cutoff score. One traditional measure of utility or value to an organization of selecting "quality" applicants on the basis of tests is simply expressed in correlational terms--the validity coefficient.

Cronbach and Gleser (1965) call this type of selection an institutional decision in contrast to an individual decision. The decisionmaker is attempting to maximize organizational benefits from a large number of similar decisions over time. Since each decision involves the same set of values, e.g., productivity or tenure, the decisionmaker can combine individual decisions statistically and obtain the best overall outcome. Chief concern is clearly with the policy or strategy of using a selection program and test results for the benefit of the organization rather than merely for the decision to accept or reject an individual for a job.

Over the last fifty years, new and more complex decisionmaking models than those based solely on regression analyses have been developed to evaluate personnel selection strategies. At first the need for different models grew out of concern for more realistic evaluation of selection programs in the specific organizational context in which the program was to be used. Today there is increasing need for evaluating selection programs in economic terms--e.g., expected costs versus benefits--because of rising personnel costs and the impact of personnel on productivity. From an organizational perspective, there appears to be a consensus that making and evaluating human resource decisions should be based on the same procedures and standards applied to all other organizational decisions.

Increasingly sophisticated decision-theoretic models to meet these new demands expressed utility in a variety of ways:

- increases in the number of successful employees (Taylor & Russell, 1939);
- increases in the average level of performance of selected employees (Brogden, 1946);
- increases in dollar-valued performance (Brogden, 1949; Cronbach & Gleser, 1965); and
- increases in average criterion scores for given validities and selection ratios (Naylor & Shine, 1965).

In the 1970s another perspective emerged, one that considered selection "fairness" from the viewpoint of minority representation in the work force. From the vantage of the organization, the goal to be achieved by the elimination of bias still is to hire the highest "quality" individuals as a means of increasing productivity; but from the viewpoint of "fairness" advocates, the goal is compensatory hiring or equality of selection outcome for various subgroups.

Proponents of "culture-fair" selection have suggested a number of models based on competing definitions of "fairness" including Cole (1973), Darlington (1971), Einhorn and Bass, (1971), and Thorndike (1971). While the various models take into account test validities, success rates of subgroups and selection ratios, each model results in different outcomes because of differing sets of implicit value judgments embedded in the selection strategies employed. Clearly none of these subgroup parity models meets the organizational view of selection fairness.

More recently to be proposed or developed have been decision-theoretic selection "fairness" models able to take into account selection situations varying in test validities, subgroup predictor and criterion scores and subgroup outcomes (Cronbach, Yalow & Schaeffer, 1980; Dunnette & Borman, 1979; Gross & Su, 1975; Petersen & Novick, 1976; Sawyer, Cole & Cole, 1976). The distinctive feature of these models is that they elicit explicit value judgments for each possible outcome and consequently permit a more rigorous evaluation of trade-offs and policies. Additionally, these utility models, providing quantitative estimates based on both validities and outcome values, can be more readily examined and publicly debated.

D. THE DECISIONMAKING PROCESS

The development of models for making more coherent decisions begins with establishing goals and objectives contributing to organizational effectiveness. A problem may arise when there is a disparity between goals and objectives and the actual results achieved. An essential step is to specify problems and then generate feasible alternatives (potential solutions) to solve them.

In utility theory, the analysis of the overall problem is accomplished by essentially breaking the problem into smaller problems that can be solved separately and then combined to provide a solution for the larger problem. A problem is decomposed into a number of decisions (alternatives) and a number of uncertain events, each with a designated probability of occurrence. The combination of each decision outcome with each event results in a foreseeable consequence. A decision table is developed that contains the probability of uncertain events in columns and numerical values (utilities) associated with the consequences of the decision in rows. Since numbers are associated with the events and decisions, it is possible to combine probabilities with utilities by an additive linear combination rule and obtain an expected utility for each decision or alternative. Each alternative is evaluated and compared with others so that the alternative selected will provide the most favorable net outcome. The best alternative is the one with the highest expected utility. Thus the choice of an alternative from among those considered is greatly dependent on the judgment of perceived values associated with consequences of decisions.

Because the decisionmaker rarely considers all possible alternatives and events or makes precise estimates of probabilities, the selected alternative rarely represents an optimal solution. The perceived expected value of the alternative chosen is only higher than the perceived expected values of other alternatives considered.

In order to construct a decision table, the decisionmaker must make his views of the decision problem explicit and follow formal procedures that permit him to consider a large amount of information in a systematic manner. The application of utility theory to a problem serves as an aid in describing, analyzing, evaluating and predicting decisions. Such decisions are more likely to maximize outcomes and be more readily communicated, understood and accepted than less systematic procedures.

Choosing a personnel selection system may be characterized within the context of utility theory as a static risky decision (Edwards, 1966; Von Neumann & Morgenstern, 1947; Wald, 1950). The rational choice of a selection system from among a number of

alternative systems is based on value judgments by organizational decisionmakers and estimates of attributes that define outcomes. Determination of selection utility from study to study tends to be very similar since nearly all studies follow the prescriptions of Brogden's model of 1949: the decisions are selection procedures; the attributes are validities, effects of the selection program, testing costs, and numbers of employees; values are expressed as dollar-based performance variance; and expected utility is based on an additive linear combination rule applied to the decision table or payoff matrix.

E. UTILITY ANALYSIS

The central practical question in personnel selection is the value or effectiveness of a selection program for a given organizational use. The way selection is evaluated in recent utility models is quite different from the traditional approach. The traditional approach is primarily concerned with validity--achieving the best possible prediction of job performance, regardless of how the test battery is used, to make selection decisions. The focus of the traditional approach is on selection efficiency expressed by the validity coefficient. The higher the validity coefficient, the smaller the error in predicting actual job performance scores.

The traditional model, with its emphasis on prediction and measurement, not only ignores the kinds of decisions required in a given selection program application, but also employs an inappropriate measure of utility since all selection decisions are considered of equal value. Additionally, the traditional model fails to consider the context or external situational parameters such as the proportion of applicants selected, the standard deviation of the value of job performance, and costs.

Cronbach and Gleser (1965) comment on the one unvarying description of selection in classical treatments:

> The traditional theory views the test as a measuring instrument intended to assign accurate numerical values to some quantitative attribute of the individual. It therefore stresses, as the prime value, precision of measurement and estimation. The roots of this theory lie in surveying and astronomy, where quantitative determinations are the chief aim. . . . In pure science it is reasonable to regard the value of a measurement as proportional to its ability to reduce uncertainty about the true value of some quantity. The mean square error is a useful index of measuring power. There is little basis for contending that one error is more serious than another of equal magnitude.
>
> In practical testing, however, a quantitative estimate is not the real desideratum. A choice between two or more discrete treatments must be

made. The tester is to allocate each person to the proper category, and accuracy of measurement is valuable only insofar as it aids in this qualitative decision. (pp. 135-136)

Cronbach and Gleser point out that decision theory is more appropriate when the test battery is used in a restricted context since the context influences the evaluation of the battery. Thus the nature of every decision must be specified and the specifications dictate the appropriate model.

Brogden (1949) had reached the same conclusion earlier. On the basis of curves he developed, showing the relation of validity, selection ratio and cost of testing to savings resulting from use of selection, he questioned the appropriateness of traditional validity estimates of utility:

> Even superficial examination of the literature on selection work indicates that the generalizations derived [in his paper] have often been disregarded. A validity coefficient of 0.3 seems in practice the lowest that psychologists will accept for a test or battery of tests to be recommended for use. If the statistical interpretation of validity coefficients current in the psychological literature were to be seriously considered in practice and coefficients such as $E = \{[100 - 100\ (1 - r^{2})^{1/2}]\}$ were to be employed in evaluating efficiency of selection, tests would be required to have a validity coefficient of 0.60 or more. (pp. 179-180)

Hull (1928) appears to be the first to comment, (in a brief footnote in his book, *Aptitude Testing*), on the distinction between individual and organizational selection and the desirability of setting a high predictor cutoff score as a means of raising average performance of those selected.

Taylor and Russell (1939), acknowledging Hull's lead, recognized that the utility of a selection device varies as a function of validity and two situational factors. They developed tables of success rates that incorporate the validity coefficient, the selection ratio (the proportion of applicants hired), and the base rate (the percentage of applicants that would be successful on the job without the use of a selection device).

Brogden (1949) proposed an improved utility model that replaced the artificial dichotomy, successful or unsuccessful job performance, called for in the Taylor-Russell model with a continuous measure of job performance and adds two very significant new situational variables: the variance of dollar-valued performance to the organization and testing costs associated with selection.

With the development of Brogden's model the theoretical framework for realistic utility analyses of personnel decisions was in place, but the difficulty of obtaining cost

accounting estimates needed to transform job performance into a dollar scale hampered the use of the model. Nearly four decades later new techniques for estimating the value of products and service produced by employees were introduced (Cascio & Ramos, 1986; Cascio & Silbey, 1979; Hunter & Schmidt, 1982; Schmidt, et al., 1979). These techniques permitted estimates of dollar-valued performance to be obtained more readily and helped to open the way for decision-theoretic studies of personnel options.

In recent years, utility analysis has become the term applied to a class of decision-theoretic models that examines performance-related consequences of selection and other behavioral interventions designed to improve productivity. Cascio (1987a) states that utility analysis is an especially valuable tool in the business setting because it forces the decisionmaker to take into account the costs and benefits of decisions:

> Utility analysis is the determination of institutional gain or loss (outcomes) anticipated from various courses of action. When faced with a choice among strategies, management must choose the strategy that maximizes the expected utility for the organization across all possible outcomes (Brealey & Myers, 1984). To make the choice, management must be able to estimate the utilities associated with various outcomes. Estimating utilities traditionally has been the Achilles heel of decision theory (Cronbach & Gleser, 1965), but a less acute problem in business settings. Although difficult to calculate, institutional gains and losses may be estimated by relatively objective behavioral or cost-accounting procedures, that is, in terms of dollars. (pp. 147-148)

Boudreau and Berger (1985) categorize a group of utility models that are concerned with the composition of the work force as external employee movement programs. Such programs include employee acquisition, separations, and various combinations of both. They suggest that all utility analysis models applied to employee movements are comprised of three basic types of variables: the number of new employees or movers selected; the expected increase in productivity produced by the selected employees; and the costs of developing and applying the predictor that leads to the productivity increase.

A second category of utility models, considered in detail in a later chapter, involves organizational interventions designed to change the productivity of employees by changing their work behavior. Examples of potential benefits from interventions include training (Schmidt, Hunter & Pearlman, 1982) and performance feedback and goal setting (Landy, Farr & Jacobs, 1982). Mathieu and Leonard (1987) demonstrated the utility of a supervising skills training program in a bank. In such utility analysis applications, Brogden's equation is modified by the use of a value representing the average difference in mean performance between treated and untreated employees in place of values for the

validity coefficient and the average predictor standard score (Schmidt, Hunter & Pearlman, 1982).

As utility analysis models become more realistic and comprehensive they may not only be able to combine external employee movement as proposed in the models of Boudreau and Berger (1985), but they may also be able to combine employee movement models with other types of interventions that improve employee productivity. For example, the military services raise or lower selection standards from time to time depending mainly on the "attractiveness" of service to potential recruits. In evaluating such a change, the impact of selection on productivity should be by no means the only consideration since selection standards directly impact on virtually all other facets of the personnel system--compensation, recruiting, training and retirement costs, retention rates and promotability. If the military were able to define and quantitatively measure its personnel options from a more realistic and comprehensive system-wide perspective, it would be able to achieve much greater net benefits. Currently, neither the military nor, for that matter, any other organization, treats personnel options from such a perspective. The development and use of more comprehensive utility analysis models in simulations would be invaluable as decision aids for planning and evaluating personnel options in a systems framework.

As described earlier, the principal concern of utility analysis is the determination of value or payoff resulting from a program's consequences. Two different conceptual approaches, traditional and decision-theoretic, have been taken to measure payoff. The next chapter will provide descriptions of historical selection utility models and their associated payoff measures.

CHAPTER 2. UTILITY MODELS

The answer to the question:, "What is the value of a predictor of a given validity for a given job?" has changed greatly since World War II. From the inception of standardized selection programs, the usefulness of a predictor, according to the traditional or classical utility approach, is determined solely in terms of measurement accuracy and predictive efficiency based on the linear regression model. Increasingly, during the last four decades, the usefulness of a predictor, according to the decision-theoretic utility approach, is measured on the basis of valued organizational outcomes.

Dreher and Sackett (1983) identify three major conceptual frameworks in tracing utility models over the years:

(1) A predictor is evaluated in terms of how well it predicts a given job-related criterion for each individual tested.
(2) A predictor is evaluated not in terms of individual prediction but rather as the extent to which it improves the proportion of applicants selected who will be successful on the job.
(3) A predictor is evaluated in terms of the value to the organization of the selection strategy used, as opposed to the value of using other strategies with the same or with other predictors.

> The shift from the first to the second represented a change in viewpoint from individual to institutional decisions; the shift from the second to the third was a shift from an emphasis on the number of successes and failures resulting from a selection procedure to emphasis on the "payoff" resulting from the adoption of the selection procedure. (p. 79)

This section describes the major selection utility models associated with each framework and evaluates them from a decision-theoretic viewpoint. We rely on reviews provided by Boudreau (1988), Cascio (1982, 19871), and Hunter and Schmidt (1982).

A. UTILITY AS A FUNCTION OF VALIDITY

1. Traditional Approaches

The extent to which a predictor measure is related to a job performance measure is indicated by the validity coefficient, r_{xy}. The statistical technique used to determine the validity coefficient usually is based on the general linear model, $y = a + bx$. The use of the

validity coefficient value as a direct index of utility has been common throughout the history of selection. A number of psychometric interpretations based on simple computational operations performed on the validity coefficient have been suggested by classical measurement proponents.

Kelley (1923) developed a utility measure based on the degree to which the correlation coefficient reduced the standard error of estimate, $\sigma \text{ est} = \sigma y \sqrt{1 - r_{xy}^2}$. The formula provides the standard deviation of the errors of prediction when a predictor, x, is used to predict a criterion, y. It shows the proportionate reduction in the standard error of criterion scores predicted by the test as compared to the standard error of criterion score predicted only by the group mean. As the correlation coefficient becomes larger, the error in predicting a criterion score with a predictor is reduced. The reduction of errors is an inverse function of $\sqrt{1 - r_{xy}^2}$, a term Kelley called the coefficient of alienation.

Using the coefficient of alienation as the measure of utility, it is clear that only very high validity coefficients result in significant gains in utility. For example, a validity of 0.86 is needed to reduce the standard error by 50%.

Hull (1928) defined utility as the Index of Forecasting Efficiency (E), $E = 1 - \sqrt{1 - r_{xy}^2}$. Thus, E is equal to the percent of perfect forecasting efficiency of a predictor in predicting a criterion. As is readily noted, E is the same as one minus Kelley's coefficient of alienation. In terms of E, the efficiency of a validity coefficient of 0.50 is only 13% better than chance.

Dreher and Sackett (1983) note Hull's comment that most validity coefficients fall in the "zone of low forecasting efficiency." Hull (1928) states "the sooner these facts are fully realized, the better for all" (p. 275). Dreher and Sackett point to Ghiselli's (1973) finding that the mean validity of predicting job performance for all types of tests is 0.22 (confirmed more recently in Schmitt, Gooding, Noe & Kirsch, 1984, survey giving a validity of 0.28 across all tests); if, then, E were to be accepted as an appropriate utility index, personnel testing is indeed in trouble.

During the 1930s and 1940s, the Index of Forecasting Efficiency was followed in popular acceptance by the coefficient of determination. This latter index is simply the square of the validity coefficient, r_{xy}^2, and indicates the proportion of job performance accounted for by the predictor. A predictor correlating 0.50 with a criterion accounts for 25% of the variance in the criterion.

When these traditional approaches are evaluated from a decisionmaking perspective, they are found to have no direct bearing on the value of a selection program or the economic benefits derived from selection. They do not relate the validity coefficient to the increase in performance of those selected by the test as compared to those selected by chance. Also, these early utility models do not properly consider the external context of the selection situation nor make appropriate assumptions concerning the utility function.

In a realistic selection situation, utility depends not only on the validity of the predictor but on the selection ratio (the ratio of the number selected to the number of applicants considered), and either the base rate (proportion of employees considered successful under current procedures) or the variance of the value of employees to the organization (spread between exceptional and poorly performing employees). All three parameters should be incorporated in a utility measure for decisionmaking. Additionally, the number of applicants to be selected, the length of their service in the organization and the costs associated with implementing and maintaining a selection program should be incorporated in a model that reflects the net benefit of a selection situation.

Traditional selection utility indices are based on combining measures of squared deviations from a predicted linear function. Any deviation of the predicted criterion value is considered equally undesirable at all points in the predictor-criterion space. Thus, concern in the traditional validity approach is maximizing correct hires and rejections while minimizing erroneous hires and rejections. Both types of errors decrease as validity increases, but the traditional approach implicitly treats all possible selection-decision outcomes as equally good or bad. One could reasonably argue that the important deviations from predictions are those that result in erroneous hires or erroneous rejections. But, in fact, in practical selection situations, organizations attach different utilities to different selection-decision outcomes. Many business organizations are totally unconcerned with erroneous rejections. However, military organizations are greatly concerned with erroneous rejections because of very sizeable recruiting costs. Airlines are greatly concerned with erroneous selection of pilots, and are not nearly as concerned with erroneous selection of baggage handlers. Again, an employer who reduces one type of error necessarily reduces the other type of error, but the type of error or outcome may be valued differently.

2. Validity as a Direct Index of Selective Efficiency

Brogden (1946a) showed that the validity coefficient itself is a direct index of selective efficiency. But Curtis and Alf (1969) note that even after more than 20 years had elapsed, Brogden's paper had not received the attention it merited--most personnel selection textbooks of the time did not even mention this important finding. In recent years, however, a marked change is evident (Cascio, 1987; Dreher & Sackett, 1983; Hunter & Schmidt, 1982; Schmidt et al., 1979). Landy et al. (1982) write:

> Perhaps the most important single outcome of the Brogden-Cronbach-Gleser approach to utility was the finding, first reported by Brogden (1946), that the validity coefficient of a selection device is the proportion of maximum utility which is attained for particular conditions of the selection ratio and the standard deviation of performance. Maximum utility is defined as the productivity gain that would occur with a perfectly valid selection device (assuming negligible selection costs). (p. 17)

When the predictor and criterion are continuous and identical in distribution form, the regression of the criterion on the predictor is linear, and the selection ratio is held constant, a predictor with a validity of 0.50 would produce 50% of the gain resulting from a perfect selection device, i.e., the criterion. Thus if employees could be selected on the basis of an actual job performance measure, and this would save an organization $300,000 per year over random selecting, then a selection device with a validity of 0.50 could be expected to save $150,000 per year (Schmidt & Hoffman, 1973).

Brogden (1946a) showed that r_{xy} could be expressed as the ratio of the mean performance of those selected using the predictor to the mean performance of those selected if selection were done on the basis of the criterion. The validity coefficient is expressed as the ratio:

$$r_{xy} = \frac{\bar{Z}_{y(x)} - \bar{Z}_{y(r)}}{\bar{Z}_{y(y)} - \bar{Z}_{y(r)}} \qquad (2.1)$$

where

$\bar{Z}_{y(x)}$ = the mean job performance (y) standard score for those selected using the test (x);

$\bar{Z}_{y(y)}$ = the mean job performance standard score resulting if selection were on the criterion itself, at the same selection ratio;

$\bar{Z}_{y(r)}$ = the mean job performance standard score resulting if selection decisions were made randomly (from along the otherwise screened pool of applicants);

r_{xy} = the validity coefficient.

Cross-multiplying Equation (1) gives:

$$r_{xy}\ (\bar{Z}_{y(y)} - \bar{Z}_{y(r)}) = \bar{Z}_{y(x)} - \bar{Z}_{y(r)} \quad .$$

Since $\bar{Z}_{y(y)} - \bar{Z}_{y(r)}$ is a constant when the proportion selected is held constant, $\bar{Z}_{y(x)} - \bar{Z}_{y(r)}$ is a linear function of r_{xy}.

Curtis and Alf (1969) compared the validity coefficient, r_{xy}, against several measures of practical significance including the increase of the criterion mean, and the proportion of "satisfactory" employees. They found r_{xy} to be a linear function of the increased criterion mean and very nearly a linear function of the other measures of practical significance.

Schmidt and Hoffman (1973) compared actual savings resulting from use of a selection device to savings predicted from Brogden's interpretation of the validity coefficient, r_{xy}; the proportion of "satisfactory" employees; and the general utility equation used in selection decisions (to be described later in this section). Cost accounting techniques were employed to estimate the loss incurred as a result of turnover in a nurse's aide job. They found that the three models provided equal accuracy of savings estimates and the fit of the three models to the cost accounting estimates were close even when violations of statistical assumptions occurred.

Hunter and Schmidt (1982) state that Brogden's formulation as expressed in Equation (2.1) has implications for the development of new techniques for directly estimating validity from reasonably accurate estimates of $\bar{Z}_{y(x)}$ and $\bar{Z}_{y(y)}$. While not using Equation (2.1) explicitly, Schmidt, Hunter, Croll and McKenzie (1983) showed that useful estimates of the validity of cognitive tests can be made by expert judges referring to validities obtained from large sample sizes. Hirsh, Schmidt and Hunter (1986) suggested that even less experienced judges could provide more accurate validity estimates than validity estimates actually computed using small samples.

Brogden's formulation shows that r equals the proportion improvement over chance that is possible with each selection ratio and that r is a linear function of the difference between the criterion means of the selected group and the population. While this interpretation of the validity coefficient represents the most accurate portrayal of the use of a predictor as an aid in the decision process, the validity coefficient alone does not provide a complete index of the effect of a predictor on the quality of decisions; factors already

mentioned unrelated to the validity of the predictor such as the selection ratio and variance of job performance also must be considered.

Brogden's (1946) interpretation of *r* was not advanced as an estimate of utility, but a later embellished version of his model (1949) defined utility as the net gain in dollars attributable to selection after testing costs were subtracted. This extended utility model will be discussed later in this section.

B. UTILITY AS A FUNCTION OF THE SUCCESS RATIO

The traditional psychometric interpretation prevailed until the development of the now widely known Taylor and Russell (1939) model. Dreher and Sackett (1983) note that at the heart of the Taylor-Russell approach is a move away from individual prediction to institutional decisionmaking along with a different perspective of what constitutes an error of prediction.

The Taylor-Russell approach reformulates the concept of measurement accuracy to that of accuracy of predicting decision outcomes. The organizational goal becomes that of making the largest number of correct decisions--identifying potentially "successful" and "unsuccessful" employees in the applicant group. The focus is no longer on the error of measurement associated with the predicted performance score of each applicant, but on the aggregate number of "true positives" (applicants predicted to succeed and who do) and "true negatives" (applicants predicted to be unsuccessful and who actually would have been unsuccessful if accepted).

In evaluating outcomes, the Taylor-Russell approach goes beyond the validity coefficient to consider two additional external parameters: the selection ratio (the proportion of applicants hired); and the base rate (the percentage of applicants who would be "successful" or "satisfactory" without the use of the selection procedure).

The Taylor-Russell model applies to a situation where employees can be conceptually placed in one of four categories: selected and successful; selected but unsuccessful; rejected but would have been successful; and rejected and would have been unsuccessful. Curtis (1967) suggested alternative terms for the four decision-outcome combinations: correct acceptances (persons accepted who succeeded); correct rejections (persons rejected who would have failed if given the opportunity); erroneous rejections (persons rejected who would have succeeded if given the opportunity); and erroneous aceptances (persons accepted who failed). Thus the model can provide outcomes for a

variety of situations such as giving the percent increase in successful employees selected by means of a new selection procedure with a given cutoff score.

Taylor and Russell developed extensive tables to determine success rates of new selection procedures that would result from various combinations of validity coefficients, selection ratios and base rates. The tables are based on the assumptions of bivariate normal, linear, homoscedastic relationships between predictor and criterion. The tables show that even predictors with relatively low validities can greatly increase the percentage of successful among those selected when the selection ratio is low. For example, when the base rate is 50% and the selection ratio is 0.10, a predictor with a validity coefficient of 0.20 will increase the percentage among the selectees who are successful from 50 to 64%--a gain of 14 additional employees per hundred selected. The tables also show that when other variables are held constant, the success rate is higher when validities are higher, selection ratios are lower, and base rates are closer to 0.50.

Hunter and Schmidt (1982), while acknowledging the advantages of the Taylor-Russell approach, describe a number of disadvantages. According to them, the foremost limitation is the need to use a dichotomous criterion and thus ignore real differences in performance (Cronbach & Gleser, 1965). All employees within a group are assumed equal in value, even though, for example, exceptional employees in the successful group might be producing two or three times above the minimum standard and other employees in the successful group might be producing at a minimally acceptable level. This method of aggregating also makes it difficult to express utility in units that are comparable across situations.

A second disadvantage of the Taylor-Russell approach is that the decision as to where to draw the line between successful and unsuccessful employees to create the dichotomy in job performance is often arbitrary. Objective information needed for deciding where to draw the line is rarely available. Decisionmakers using different standards may define the dividing line quite differently. Since the value of a selection procedure depends on the base rate of success, the model could lead to quite different outcomes depending on where the line is drawn.

In addition to the use of the Taylor-Russell model in situations in which the criterion is dichotomous, e.g., occurrence of turnover, the model may be appropriate for a number of other situations as suggested by Cascio (1987a). Examples given by Cascio are jobs in which differences in performance beyond the minimum do not change benefits

(clerical, technicians) and jobs in which output differences are unmeasurable (nursing, teaching, counseling). However, Cascio does not provide data supporting the existence of such types of jobs.

The Taylor-Russell model was the first to show the complex interactions among the validity coefficient, the selection ratio and the base rate. The way these interactions affect decision outcomes in the model makes it abundantly clear that the context of a selection situation must be specified to evaluate its value.

Utility, however, expressed in terms of increases in the success ratio, does not reflect how much real difference there is in performance among employees. The model also is considered to lack general applicability because of the somewhat arbitrary decision involved in creating the job dichotomy. Further, as a decisionmaking model, it does not consider the number of employees selected or costs.

Sands (1973a) developed a means of attaching cost of attaining personnel requirements (CAPER) to decision outcomes. The purpose of the model is to determine an optimal recruiting-selection strategy that minimizes the total cost of recruiting, selecting, inducting, and training individuals to meet a quota of satisfactory personnel. The model requires the computation of graduates and failures separately that would qualify for acceptance at each possible alternative cutting score on a selection test. While an advantage of the model is that results can be given in dollar terms, a disadvantage is that it uses, like the Taylor-Russell model, a dichotomous criterion. A later simplified version developed by Sands (1973b) assumes a bivariate normal distribution of test and criterion scores.

C. UTILITY AS A FUNCTION OF INCREASE IN THE CRITERION SCORE

Brogden's (1946a) formulations, described earlier, eliminated both the problem of forcing a dichotomized criterion and using a two-point distribution of job performance rather than the full range of variations in job performance. Brogden's index of utility is simply the increase in a continuous criterion score. Brogden showed, based on the general linear prediction equation, that the best predictor of the average standard criterion score can be expressed as:

$$\bar{Z}\,y = (r_{xy})\,(\bar{Z}x) \quad . \tag{2.2}$$

where

r_{xy} = the validity coefficient;

$\bar{Z}_x$ = the mean test standard score for those selected using the test;

$\bar{Z}_y$ = the mean job performance standard score for those selected using the test.

The values of r_{xy} and $\bar{Z}_x$ can be computed directly from the selection situation. Kelley (1923) showed that when predictor scores are normally distributed and the predictor-criterion relationship is linear, then $\bar{Z}_x$ may be computed by the formula λ/ϕ, where λ is the height of the normal curve at the point of cut, and ϕ is the percentage in the selected group (the selection ratio). Thus Equation (2.2) may be expressed as:

$$\bar{Z}_y = r_{xy}\ \lambda/\phi \quad . \tag{2.3}$$

The use of Kelley's equation further assumes that applicants are rank-ordered by predictor score and are then selected from the top down.

Brogden's formulations show that given any arbitrarily specified cutoff on a predictor, the higher the validity, the greater the increase in mean criterion score for the selected group over the score of the total applicant group.

Naylor and Shine (1965), employing Equation (2.3), developed an extensive series of tables that readily permit the computing of mean criterion scores for the selected group. For each selection ratio, the corresponding standard predictor cutoff value, the ordinate of the normal curve, and the mean standard predictor score are given.

The increase in mean criterion performance is a practical utility index that is more generally applicable in selection programs than the Taylor-Russell approach. Because the increase is expressed in standard score units, meaningful comparisons across studies using different criteria also can be made. Although both models employ validity coefficients and selection ratios, the significance of both components are more readily apparent.

One significant limitation of increases in criterion performance as an index of utility is that the index is expressed in standard score units rather than in terms of dollars or production quantities that decisionmakers are accustomed to weigh in making evaluations. However, Brogden's later formulations in 1949 expanded the components of the equation to explicitly consider payoff in dollar terms, costs and other external parameters of the selection situation. These formulations are described in the next section.

D. UTILITY AS A FUNCTION OF DOLLAR-VALUED PERFORMANCE

1. Development of Models

As indicated earlier, Brogden (1946a) showed that the validity coefficient can be interpreted as the ratio of the saving actually achieved by a selection device to the saving actually attained by selection on the criterion itself. "Savings" is defined as the difference between the mean criterion score of the selected group and the mean of the applicant population. Brodgen (1949) developed the concept of savings or economic utility further:

> If the criterion can be expressed in cost-accounting terms as an estimate of the dollar saving effected by selecting the given individual instead of an average applicant, the coefficient, $r\,\sigma_y$, estimates the expected or average dollar saving achieved by a unit increase in the standard predictor score. Moreover, if M_{sx} denotes the average standard test score of the selected group, $r\,\sigma_y\,M_{zx}$ gives the mean gain in production or, if the criterion is expressed in dollars, the average or expected saving from selection. . . . The validity coefficient, r_{xy}, gives the percentage of possible saving; the product, $r_{xy}\,\sigma y$, estimates the increase in saving per unit increase in average predictor score, and the last term $r_{xy}\,\sigma y\,M_{zx}$ gives the actual average saving for a given test and a given selected group. (p. 178)

In other words, Brogden showed that utility increases with increases in the validity coefficient, the standard deviation of dollar-valued performance, and the mean predictor score of those selected. Selection costs, however, increase with decreases in selection ratios and are subtracted from productivity gains to estimate total net economic benefit.

Brogden's (1949) derivations have been detailed more recently by Cascio (1982, 1987a), Hunter and Schmidt (1982), and Schmidt et al. (1979). Brogden's landmark equation requires only the assumption of linearity between predictor and job performance. The elaboration of Brogden's equation below follows Hunter and Schmidt's (1982) description.

Let r_{xy} equal the correlation between the predictor (x) and job performance measured in dollars (y):

$$Y = \beta Z_x + \mu_y + e$$

where:

Y = job performance measured in dollar value;

β = the linear regression weight on test scores for predicting job performance;

Z_x = test performance in standard score form *in the applicant group*;

μ_y = mean job performance (in dollars) of randomly selected employees;

and

e = error of prediction.

For those selected (s), the equation that gives the average performance is:

$$E(Y_S) = E(\beta Z_{xs}) + E(\mu_y) + E(e)$$

Since the expected value E(e) is zero, and β and μ are constants, the equation simplifies to:

$$\overline{Y} = \beta \overline{Z}_{xs} + \mu_y \quad .$$

Noting that $\beta = r_{xy}\,(SD_y / SD_X)$ and SD_y is the standard deviation of dollar-valued performance among randomly selected employees, the equation can be further reduced. Also, since $SD_X = 1.00$, $\beta = r_{xy}\,SD_y$, the equation becomes:

$$\overline{Y} = r_{xy}\,SD_y\,\overline{Z}_{xs} + \mu_y \quad .$$

The equation gives the *absolute* dollar value of average job performance in the selected group. What is wanted is an equation that gives the *increase* in dollar value of average performance that results from using the predictor. To obtain such an equation, we first note that $\overline{Y}_s = \mu_y$ if the predictor were not used, i.e., mean performance in the selected group is the same as mean performance in a group selected randomly from the applicant population. The increase due to the use of a valid predictor, then, is $r_{xy}SD_y\overline{Z}_{xs}$. The desired equation is obtained by transposing and is:

$$\overline{Y}_s - \mu_y = r_{xy}\,SD_y\,\overline{Z}_{xs} \quad .$$

The right side of the above equation represents the difference between mean productivity in the group selected using the predictor and mean productivity in a group selected randomly (without using the predictor). The equation that gives mean gain in productivity per selectee (marginal utility) resulting from the use of the predictor is:

$$\Delta\overline{U}\ /\text{selectee} = r_{xy}\,SD_y\,\overline{Z}_{xs} \qquad (2.4)$$

where U is utility and ΔU is marginal utility.

Equation (2.4) states that the average productivity gain in dollars per person selected is the product of the validity coefficient, the standard deviation of job performance in dollars, and the average standard score on the predictor of those selected. The value $r_{xy}\,\overline{Z}_{xs}$ is the mean standard score on the dollar criterion of those selected, $\overline{Z}_y$. Utility

per selectee is the mean Z score on the criterion of those selected times the standard deviation of the criterion in dollars.

The total utility of the predictor depends on the number of persons selected or hired. The total utility or productivity gain is the mean gain per selectee times the number of people selected, N. Thus the total productivity gain is:

$$\Delta\overline{U} = N_s\, r_{xy}\, SD_y\, \overline{Z}_{xs} \quad .$$

Schmidt et al. (1979) included a term in the equation for expected tenure, T, of one selected cohort group. Additionally, until this point in the formulation, the cost of testing has not been considered. Although in typical situations testing costs are small, especially in relation to the value of utility gains, Brogden's equation included testing costs. When including length of tenure and cost, Equation (2.4) becomes:

$$\Delta U = TNr_{xy}SD_y\,\overline{Z}_{xs} - NC \qquad (2.5)$$

where ΔU = the total utility gain of a selection program;

T = the mean number of years selectees remain on the job;

N = the number of individuals selected;

r_{xy} = the correlation between the predictor and job performance in the applicant population (the validity of the predictor);

SD_y = the standard deviation of job performance in dollars;

$\overline{Z}_{xs}$ = the average predictor score of those selected in applicant population predictor standard score;

C = the cost of selection per individual.

The values for r_{xy} and SD_y should be those that would pertain to the applicant population, the group on whom the selection procedure is actually to be used. Obviously, values used with incumbents would be underestimates because of range restriction on both predictor and job performance measures. Hunter and Schmidt (1982) suggest obtaining estimates of true validity by employing corrections for both restriction in range and criterion unreliability.

Both Brogden (1949) and Cronbach and Gleser (1965) noted that if it is assumed that test scores are normally distributed, a more restrictive requirement than Brogden's Equations (2.4) and (2.5) requirement of only linearity in the predictor-criterion relationship, an alternative formula can be used for derivational or computational convenience. The mean test score of those selected, $\overline{Z}_{xs}$, in Equations (2.4) and (2.5) is

replaced by ϕ/p where ϕ is the ordinate of the normal distribution at the cutoff score and p is the selection ratio. By using a table of area of the normal curve which gives ordinate values, it sometimes may be more convenient to obtain ϕ/p than to compute $\bar{Z}_{xs}$. When ϕ/p is substituted for $\bar{Z}_{xs}$ in Equation (2.4), the mean gain in productivity per selectee resulting from the use of the test is:

$$\Delta U/\text{selectee} = r_{xy}(\phi/p)SD_y \quad (2.6)$$

In their highly influential book, *Psychological Tests and Personnel Decisions*, Cronbach and Gleser (1965) note that formulas they developed for traditional selection utility, single-stage fixed job selection decisions, are identical to those developed by Brogden (1946, 1949). Cronbach and Gleser's equations are initially derived in terms of mean gain per applicant; Brogden's equations are derived in terms of mean gain in utility per selectee. However, Brogden's equations are readily shown to be equivalent to the Cronbach and Gleser (1965) equation for total utility. Collectively the equations are often referred to as the Brogden-Cronbach-Gleser selection utility model. Cronbach and Gleser (1965) also made original contributions in the development of new utility concepts for applications involving placement and selection procedures using sequential decisions and multivariate data.

Hunter and Schmidt (1982) note that a glance at Equations (2.4) and (2.6) shows that the validity coefficient enters the equation as a multiplicative factor. Thus, increasing or decreasing the validity by any factor will increase or decrease the utility by the same factor. If validity is raised from 0.20 to 0.40, Equations (2.4) and (2.6) show that utility doubles. Equations (2.4) and (2.6) also show the limitations on utility, even for a predictor of perfect validity. A very high selection ratio and a very low SD_y could reduce utility gains to a level of little value.

In Equation (2.6), $(r_{xy})(SD_y)$ is the slope of the payoff function relating expected payoff to score. The slope depends on both the size of the validity and dispersion of criterion scores. An increase in validity leads to an increase in slope, but the practical significance of individual differences in payoff also varies with the magnitude of SD_y. When SD_y is large, even selection programs with low validity can be quite useful. Hunter and Schmidt (1982), employing previously described equations and, excluding duration effects and costs, provide the simple example:

	r_{xy}	$\bar{Z}_{xs}$	SD_y	ΔU/ selectee
Mid-level job (e.g., systems analyst)	0.20	1.00	25,000	$5,000
Lower-level job (e.g., janitor)	0.60	1.00	2,000	1,200

In this example, the average, marginal utilities are $5,000 and $1,200. If 10 people were hired, the actual utilities would be $50,000 and $12,000 respectively. If 1,000 people were to be hired, then the utilities would be $500,000 and $120,000, respectively. Obviously the *total* dollar value of tests is greater for large employers than for the local shoeshine stand. However, this is misleading, because on a *percentage* basis it is average gain in utility that counts; and that's what counts to each individual company. (p. 237)

2. Testing Costs

Both Brogden (1949) and Cronbach and Gleser (1965) included testing costs as an important consideration in evaluating selections decisions. Although their original models focussed on the actual cost of administering tests to applicants, testing costs properly should include all costs associated with the selection program such as incremental costs of recruiting applicants of higher ability levels and the cost of developing and implementing a new selection battery. In most business applications, the cost of administering tests is very small compared to selection utility. Brogden (1949) showed that when testing cost is very high, $\Delta\bar{U}$ /selectee will be less at very low selection ratios than at higher selection ratios. Although highly unlikely in most situations, testing costs at extremely low selection ratios can result in a loss of utility. Brogden (1949) suggested that "the ratio of cost of testing to the product of the validity coefficient and σ_y in dollar units should not exceed 0.10. It would be desirable to hold it below 0.05." (p. 177)

In order to meet manpower "quality" requirements, the military services attempt to influence the size of the applicant pool by intensive recruiting efforts, advertisements, and a variety of economic inducements. About 315,000 enlistees are recruited by the military services each year and the process involves very sizeable expenditures for obtaining and testing potential recruits.

Cronbach and Gleser (1965) provide a formula to determine the number of applicants to be tested to obtain the needed quota of new employees to maximize productivity gains from selection:

$$\phi(p) - pZ_x = C/(r_{xy}SD_y) \qquad (2.7)$$

where Z_X is the cutting score on the test in Z score form. The equation must be solved by iteration. Only one value of p (the selection ratio) will satisfy the equation and p will always be less than or equal to 0.50. The formula assumes that the cost of testing per applicant remains constant regardless of the total number tested. However, in the military context a different range of values for p would obtain, since the cost of recruiting varies with varying selection ratios, e.g., the cost of recruiting higher quality applicants is greater than the cost of recruiting lower quality applicants.

3. Advantages and Limitations of the Brogden-Cronbach-Gleser Model

The Brogden-Cronbach-Gleser model provides a more comprehensive framework for decisionmaking than all other models. Boudreau and his co-authors state that this model and other utility analysis models could be understood in terms of quantity, quality, and cost. Rich and Boudreau (1987) summarize:

> Quantity refers to the number of employees and time periods affected by the program. Quality refers to the change in average employee value (per employee, per time period) resulting from the program. The sum of the average quality change across the quantity of employees and time periods provides an estimate of the program's returns. The cost of the program reflects the resource commitments necessary to develop, implement, and continue the program over the period of analysis. Utility models estimate these three components using a variety of parameters (e.g., standardized program effects, correlation coefficients, dollar-value scaling factors), estimating an expected value for each parameter, and then mathematically combining the parameter estimates to derive an expected utility value. (p. 57)

Since the Brogden-Cronbach-Gleser model is comprised of all three components and provides an incremental utility index that permits comparisons to be made between existing and new selection programs in dollar terms, it is superior to other major models described in this section as an aid in decisionmaking.

Over the last several years there have been a number of refinements and extensions proposed for the basic Brogden-Cronbach-Gleser model. These modifications include: consideration of more than a single cohort group of new employees along with employee movement data; alternative methods of estimating SD_y in dollar terms; more sophisticated means of incorporating financial and economic factors; and alternative strategies and rules in the hiring process. Although these changes will be discussed in greater detail later in the report, they are briefly noted in this section to indicate some possible concerns about the meaning and accuracy of the Brogden-Cronbach-Gleser utility measure.

The Brogden-Cronbach-Gleser model considers only a single cohort group of employees hired on the basis of the new selection program. Usually selection program effects (productivity benefits) are multiplied by the mean tenure of the cohort group. In such practice, the model may underestimate marginal utilities because it does not take into account subsequent groups of employees hired on the basis of the new selection program, nor does it take into account employee movements (turnover and internal staffing) as proposed by Boudreau (1983b) and Boudreau and Berger (1985).

The Brogden-Cronbach-Gleser utility estimate is subject to varying interpretations in part because of the different approaches used in measuring SD_y, each approach yielding somewhat different values. Cascio (1987a) describes six major approaches in measuring SD_y: cost-accounting (Brogden & Taylor, 1950; Roche, 1961); the 40% rule (Hunter & Schmidt, 1982; Schmidt & Hunter, 1983); global estimation (Schmidt et al., 1979); CREPID (Cascio & Ramos, 1986); and system effectiveness and superior equivalent techniques (Eaton, Wing & Mitchell, 1985).

A number of researchers regard the accurate measurement of SD_y as essential to understanding and improving utility estimates of expected payoff (Bobko, Karren & Parkington, 1983; Burke & Frederick, 1984; Greer & Cascio, 1987; Reilly & Smither, 1985; Weekley, Frank, O'Connor & Peters, 1985; Simonet, 1984). The great majority of analyses conducted in the 1980s compare one method of obtaining SD_y with another, but "goodness" of method is generally judged on psychometric grounds rather than on the basis of making better selection decisions. These studies usually have not included a utility analysis.

The Brogden-Cronbach-Gleser utility estimate is subject to varying interpretations in part because of the limited economic considerations incorporated in the basic model (Boudreau, 1983a; Cronshaw & Alexander, 1985). Boudreau cautions that marginal utilities may be overestimated if the utility model does not take into account three economic factors: variable costs--costs that covary with productivity such as raw material costs, productivity overhead, and commissioned-base pay; marginal tax rates--the tax liabilities of an organization's profit associated with increased utility due to improved selection; and discount rates that adjust future costs and benefits to reflect lost opportunity costs since present money can at least be invested at prevailing interest rates. Boudreau (1983b) modified the Brogden-Cronbach-Gleser model to incorporate both employee flow and economic variables in the equations. Hunter, Schmidt and Coggin (1988) argued,

however, that "different conceptual definitions of utility are useful under different circumstances; there is no single 'correct' definition of utility." (p. 522)

The Brogden-Cronbach-Gleser model does not consider the effects of variability and risk in making utility estimates. Rich and Boudreau (1987) suggest that the size and shape of each of the components used in the utility equation be computed and that each of the component distributions be combined through a Monte Carlo analysis to yield a distribution of total utility values. Utility analysis based on uncertainty should better reflect the selection program situation than a point estimate of expected utility value.

Many utility applications give potential productivity gains based on a range of selection rates assuming a top-down hiring strategy of selection, starting with the highest scoring applicant. Further, it is assumed that all applicants offered a job will accept. It is further assumed implicitly that the organization's applicant pool is representative of the overall applicant pool.

Such assumptions project optimal selection gains that rarely correspond to realistic selection procedures. Schmidt, Mack, and Hunter (1984) examined the impact of a valid selection test on the productivity of forest rangers. A top-down selection procedure would result in a productivity gain of about 13% as compared to an increase of about two percent employing a minimum cutting score set at one standard deviation below the mean. Hunter and Hunter (1984) found that using a top-down strategy *within* both minority and non-minority applicant groups preserved 95% of the gain obtained with a top-down strategy using a single, combined group. However, using a low cutting score would decrease gains by 83%. Murphy (1986) developed equations for computing ability scores of applicants that accepted employment compared to scores of all applicants receiving offers of employment and concluded that utility formulas could overestimate gains by 30 to 80%. But Schmidt, Hunter, McKenzie and Muldrow (1979) showed that calculations of $\bar{Z}_x$, based on the normal curve can be adjusted to allow for corrections, e.g., if 10% are hired but 20% must be offered jobs to get the 10%, the SR used in the calculation is 0.20. Murphy (1986) did not consider this procedure.

If, however, the mean score, $\bar{Z}_{xs}$, of those selected in the Brogden-Cronbach-Gleser model is used to evaluate gains in actual selection programs rather than ϕ/p, no unrealistic assumptions are required or made. The only essential assumption is the universally accepted linear relationship between predictor and job performance measures.

One additional limitation of the Brogden-Cronbach-Gleser model is that it considers only change in productivity brought about by the improved quality of the workforce. Other factors such as training, reward systems, employee movement, and design of equipment are not considered although these interventions interact with the composition of the workforce and may significantly impact human resource decisions.

The Brogden-Cronbach-Gleser model, incorporating expanded equations for considering employee flow and economic factors, is judged, on the basis of recent widespread attention received in research and in applications, to be the most versatile, generally applicable and accepted decision-theoretic model for operationalizing and measuring the economic benefits of predicting job performance.

4. The Use of Selection Utility Models

Hunter and Schmidt (1982) point out that despite the availability of Brogden's equation since 1949, utility analyses did not receive widespread attention until recent years. The reasons for the sparsity of attention are not entirely clear, although various explanations have been offered including: lack of interest in economic measures; difficulty in explaining utility concepts, equations, and measurement procedures to decisionmakers; and the continuing belief that precise estimates of all the components of the equation were needed for generating an accurate expected utility value for choosing among alternatives.

Hunter and Schmidt (1982) attribute the lack of work in utility mainly to three factors. First, many psychologists believe that utility equations are of no value unless the data *exactly* fit the linear homoscedastic model and all marginal distributions are normal. Since Brogden's utility model is based on the general linear model, it shares the same statistical assumptions made for the general linear model: a linear relationship between predictor and job performance measures, equality of variance between predictor and job performance scores (the conditional distributions), and normality of the conditional distributions.

As discussed earlier, the basic selection utility equations depend only on linearity. Equation (2.6) also assumes normality of the predictor-score distribution, but this assumption essentially was introduced by Brodgen (1949) and Cronbach and Gleser (1965) for derivational convenience. Since the mean predictor score, $\overline{Z}_{xs}$, can be computed directly, the normality assumption is not required. Further, the equality of variance assumption also is not required if the conditional distributions are standardized.

Cronbach and Gleser (1965) describe a number of plausible examples that may arise in the workplace leading to job performance distributions that are positively or negatively skewed. But Hunter and Schmidt (1982) provide a review of studies that tested the linear homoscedastic model fit. In nearly all instances, the statistical assumptions were met and only occasional small departures were found. Hunter and Schmidt (1982) report on Brogden's research in the 1960s that attempted to identify nonlinear relationships in military selection studies using large samples. Brogden found that nonlinear relationships were never superior to simple linear functions in cross-validation samples, although higher-order nonlinear equations sometimes produced impressive fits in initial samples. Van Naersson (1963) also showed that violations of normal distribution for test scores produced very similar utility estimates.

Hunter and Schmidt (1982) convincingly conclude:

> Thus, it appears that an obsessive concern with statistical assumptions is not justified. This is especially true in light of the fact that for most purposes, there is no need for utility estimates to be accurate to the last dollar. Approximations are usually adequate for the kinds of decisions that these estimates are used to make (Van Naersson, 1963, p. 282; Cronbach & Gleser, 1965, p. 139). Alternatives to use of the utility equations will typically be procedures that produce larger errors, or even worse, no utility analyses at all. (p. 245)

The second reason offered by Hunter and Schmidt (1982) for sparsity of work in selection utility analyses is the widespread belief held until recent years that validity is situationally specific. Subtle differences in job performance requirements from situation to situation were thought to produce significant differences in test validities (Ghiselli, 1959, 1966, 1973). Recent research, however, supports the concept of validity generalization holding that predictor-criterion relationships are stable in similar job settings. Variability in validity coefficients across studies is due to a significant extent to artifactual differences such as differences in sample size, range restriction, and test and criteria reliabilities. Schmidt, Hunter, Pearlman, and Shane (1979) found in an analytic study that these four sources of artifactual variance accounted for an average of 62% of the variance in validity coefficients. Similar empirical results were obtained by Pearlman, Schmidt, and Hunter (1980) and in Schmidt, Gast-Rosenberg and Hunter (1980). However, Schmitt et al. (1984), employing very large sample sizes, found a higher proportion (or percentage) of remaining artifactual variance than is usually found. But McDaniel, Schmidt, Raja, and Hunter (1986) pointed out the remaining amount (as opposed to percentage) of variance

was very small, as in previous validity generalization studies. The proportion of remaining variance was attributed to the very large sample sizes used in the Schmitt et al. study.

If validity were to be accepted as situationally specific, separate utility analyses would be needed for each application. Each utility analysis then might also involve a difficult and expensive cost accounting procedure to measure dollar-valued performance. On the other hand, when validity generalization is possible, the best estimate of test validity is the mean of the corrected validity distribution. In a new situation, involving the same test type and job, Hunter and Schmidt state that "only a job analysis is necessary, in order to ensure that the job at hand is a member of the class of jobs on which the validity distribution was derived" (p. 247). Hunter and Schmidt suggest that opening the way for validity generalization findings for a wide range of test-job combinations would do much to encourage the use of utility analysis.

The final reason offered by Hunter and Schmidt (1982) is the belief that cost accounting procedures must be used in estimating the dollar value of employee performance (Brogden and Taylor, 1950; Roche, 1961; Cronbach & Gleser, 1965, pp. 254-266). Cascio (1987a) considered the difficulty in obtaining the dollar estimate the main reason for the lack of widespread attention until recent years. Roche describes a complex cost accounting procedure employed for the job of radial drill operator that involved great effort, time, expense, and reliance upon arbitrary judgments.

Hunter and Schmidt (1982) and Schmidt et al. (1979) developed a global estimation procedure for obtaining rational estimates of SD_y. The procedure requires neither cost accounting, direct measurement of performance nor job analysis. The procedure uses judgments of supervisors who have had the opportunity to observe differences in output among employees. Supervisors are asked to place a dollar value on points of an hypothetical curve of job performance. If job performance in dollar terms is normally distributed, then the difference between the value to the organization of the products and services of an employee at the 85th percentile in performance and that of an employee at the 50th percentile in performance is equal to SD_y. Supervisors are asked in making this mental judgment to consider the cost of hiring an outside consultant firm to provide the same work as their employees. Judgments of individual supervisors are averaged across a large number of supervisors to remove bias and random errors.

Hunter and Schmidt (1982) indicate that procedures similar to their global estimation techniques have been used to scale unmeasurable but critical variables in high-

level policy decisionmaking and that these procedures were well received. They reference work in such activities as constructing nuclear power plants, determining corporate risk policies, developing investment and expansion programs, and in seeding hurricanes (Howard, 1966; Howard, Matheson & North, 1972; Matheson, 1969; Raiffa, 1968). Hunter and Schmidt point out that such global estimation procedures are virtually unavoidable for higher level jobs, and that utility estimates derived from this procedure should lead to correct decisions about selection procedures.

With the reduction of concerns about the linear homoscedastic model, the demonstration of stability of validity coefficients for predictors of similar jobs, and the development of practical procedures for measuring dollar-valued job performance, the way was opened for widespread application of the Brogden-Cronbach-Gleser model of selection utility analysis. The next section of this report examines in greater detail the major scaling approaches used in translating job performance into economic terms.

CHAPTER 3. ESTIMATING DOLLAR-VALUED PERFORMANCE

A. THE PAYOFF SCALE IN DOLLAR TERMS

The utility Equation (2.4) described in the previous chapter requires the estimation of three critical parameters. Procedures for obtaining estimates of r_{xy} and $\overline{Z}_{xs}$ are relatively straightforward. The correlation of a predictor and a well-developed measure of job performance (y') provides a good estimate of r_{xy} --the correlation of the predictor with job performance measured in dollars. Hunter and Schmidt (1982) point out that job performance rating measures are often subject to ceiling effects (because of leniency in ratings) causing an underestimate of the true value of r_{xy}. Values of $r_{xy'}$, Hunter and Schmidt also note, should be corrected for both restriction in range and criterion unreliability to obtain truer estimates of correlation with job performance.

Estimates of $\overline{Z}_{xs}$ are simply obtained by computing the mean predictor scores of those selected for employment from among the applicants.

The parameter most difficult to obtain in practice is SD_y, the standard deviation in dollar-valued performance. It was once believed that estimates of SD_y could be obtained only through complex and time-consuming cost accounting procedures such as those outlined by Brogden and Taylor (1950). Such procedures involved costing the job performance of each employee and then computing the standard deviation of productivity in dollars.

During the last decade, new procedures were developed that require only the judgment of experts in estimating values to compute SD_y. While there remains only one comprehensive utility model, these new approaches offer alternative procedures that may be used for the payoff scale.

Boudreau (1988) provides the most comprehensive summary and analysis extant of SD_y measurement research. He tables results of 34 studies that produced over 100 individual estimates; five studies were accomplished between 1953 and 1978 compared to 29 studies between 1979 and 1988. Boudreau notes that the central issue examined by

research is whether different SD_y measurement techniques yield different values. Following Boudreau (1988), our chapter details methods of measuring SD_y and highlights comparisons among them.

B. METHODS OF MEASURING THE PAYOFF SCALES

1. Cost Accounting Method

Brogden and Taylor (1950) suggested the use of cost accounting procedures to develop a dollar criterion that measures the contribution of the individual to overall efficiency of the organization. They list a number of criterion elements for possible consideration:

1. Average value of production or service units.
2. Quality of objects produced or services accomplished.
3. Overhead--including rent, light, heat, cost depreciation, or rental of machines and equipment.
4. Errors, accidents, spoilage, wastage, damage to machines or equipment due to unusual wear and tear, etc.
5. Such factors as appearance, friendliness, poise, and general social effectiveness in public relations. (Here, some approximate or arbitrary value would have to be assigned by an individual or individuals having the required responsibility and background.)
6. The cost of spent time of other personnel. This would include not only the time of the supervisory personnel but also that of other workers. (p. 146)

Roche (1961; summarized in Cronbach & Gleser, 1965, pp. 254-266) conducted a field study that directly applied Brogden and Taylor's cost accounting elements to 291 beginning level radial drill operators (RDO-1s) employed by a large heavy equipment manufacturing plant. The job description for an RDO-1 is:

> Sets up and operates a radial drill, performing drill, ream, line ream, tap (stud, pipe, and standard), countersink, chamfer, bore, counterbore, spotface, backface, and hollow mill operations. Involves various types of parts such as castings, forgings, bar stock, structural steel and welded fabrications. Grinds drills when necessary. (p. 257)

Roche writes that the RDO-1s have no control over the type of parts on which they perform machine operations because the planning department assigns work by machine

number, not to specific operators. An RDO-1 typically works on a variety of parts and the mix of work will vary for each worker.

It was assumed that the dollar profit which accrues to the company as a result of an individual's work provides the best estimate of his worth to the company. The procedures used to develop this dollar criterion follow the description provided by Roche.

The company's cost accounting methods were used. The procedure is based on "standard costing" which Roche states is an effective tool for volume production accounting. It permits application of the "principle of exception" that directs attention to variations from standard cost and indicates trends in volume output. Standard product cost is based on cost data on material used to produce products, direct labor used to alter the material, and facility usage required to perform direct labor. Standard cost must remain stable for a specific period if the cost-accounting method is to work. In Roche's field study of drill operators standards usually remain frozen for a five-year period.

"Lifo" (last in-first out) inventory accounting is used, releasing the most recent inventory costs as costs of goods used or sold, thereby attempting to match the current cost of obtaining inventory against sales. Seven major cost elements are used in the accounting including material, direct labor, and general and machine burden.

Prime product costs are built up from costs of piece-parts or units into costs for assemblies, then costs for groups, arrangements, and finally costs for the general arrangement or complete model. The total cost figure has four basic components: variable, fixed, office, and parts warehousing.

Income from the RDO-1's work can be readily determined, since the parts manufactured are sold to dealers, and a price for each part has been established. Subtracting the cost at standard production from the price provides a profit picture.

A productivity measure, the "performance ratio", was obtained to express the payoff for each individual. Since standard time study procedures established the length of time for a competent operator to complete the machine operation of each piece-part, the number of piece-parts per hour that an operator should be able to process was known. An operator's performance ratio for any period of work is then computed as the actual production per hour divided by the standard hourly production for the piece-part on which the RDO-1 has been working. A performance ratio can be determined over a period of time during which the operator has worked on a number of different piece-parts, each with a different production standard. Roche states that only rarely does an operator turn out more

than standard production over an extended period of time. Performance ratios for each operator were obtained monthly for a six-month period, the mean being taken as the operator's typical performance.

Roche describes a "burden adjustment" procedure for dealing with an RDO-1 that produces at less than standard. The actual burden per hour for his inefficiency is greater than the standard burden per hour determined for his or her operation. Each below standard performance ratio was corrected by the formula, $2\text{-}1/PR$, where PR is performance ratio. If an operator is working at 80% of standard, the burden is 1.25 hours instead of 1.00 hours, and the corrected performance ratio is 0.75.

Roche outlines the procedures for determining each operator's payoff:

1. Computation of each operator's typical performance ratio. This figure was his or her mean performance ratio for the six-month period of the study.
2. Adjustment of the typical performance ratio for below-standard production (the burden adjustment).
3. Computation of the average profit at standard production, attributable to the radial drill operation.

   a. Tabulation of the standard production rate for each type of piece-part machined by radial drill operators. These data were provided by the time study division.

   b. Profit for each type of piece-part attributable to the radial drill operation.

   c. Profit per hour for each piece-part attributable to the radial drill operation at standard production. These figures were determined by multiplying the profit per piece by the standard production rate for the piece.

   d. Average hourly profit attributable to the radial drill operation at standard production. This was determined by weighting the profit per hour for each piece-part (step 3c) by the number of such parts in the work flow.
4. Determination of e [y], the profit for each radial drill operator at his or her corrected performance ratio and the standard hourly profit.
5. Computation of σe [SD_y]. This is merely the standard deviation of the e values computed in step 4. (p. 260)

Roche reported that the 291 radial drill operators worked on approximately 2,500 different piece-parts. In order to reduce the enormous amount of clerical labor to determine the profit attributable to the radial drill operators for every piece-part, a random sample of 275 parts (about 10% of the total) was used for computational purposes. Averaged over

the sample of parts, profit per hour attributable to the radial drill operators was $5.51 with a standard deviation of $3.95. The *y* value for each operator was the hourly profit at standard production attributable to RDO-1 multiplied by his or her corrected performance ratio. The standard deviation of this distribution (SD_y) was $.59. Cascio (1987a) computes this figure as $2.66 in 1987 dollars.

Roche concludes that the study clearly demonstrated that a dollar criterion, such as suggested by Brogden and Taylor, can be developed. He, however, states that, "Although the methods are relatively straightforward, many estimates and arbitrary allocations enter into the cost accounting" (p. 263). He adds that the objectivity of performance ratio figures can also be questioned and that cost accountants are not in universal agreement as to which factors should be included in performing a cost analysis.

Cronbach and Gleser (1965) comment that Roche, a psychologist, was necessarily dependent on the advice of the company's accountants and that the accountants may not have clearly perceived the problem; hopefully a more thorough interdisciplinary attack in the future will produce better solutions to the dollar payoff from an employee.

Others also have criticized the cost accounting approach in terms of complexity, difficulty, and effort entailed (Cascio, 1980; Cascio & Ramos, 1986; Hunter & Schmidt, 1982). Hunter and Schmidt called Roche's payoff scale "deficient on a logical basis" because it uses profits rather than value of goods and services; Boudreau (1983a) indicated that Roche's attribution of fixed costs to employees may be inappropriate.

Greer (1986), as reported in Cascio (1987a), attempted to use cost accounting to develop an objective estimate of SD_y for the job of route salesperson for a soft drink bottling company. He came to the same conclusion that Roche made 15 years earlier: Since cost accounting systems focus on the costs and benefits of units of product, not units of performance, many estimates are needed, and the putative objectivity of cost accounting data may become suspect.

Nevertheless, Greer and Cascio (1987) argue that a cost accounting approach to the estimation of SD_y "remains as the conceptual standard of comparison" because it is generally objective, verifiable by a third party, and subject to internal and external audit. They write:

> The cost object in Roche's (1961) study was an individual worker's performance level. For a cost-accounting system to provide a valid estimate of the cost of any cost object, it should be designed with that cost objective in mind. The cost objective and cost object were significantly different from

the cost objectives and cost objects that the cost-accounting system was designed to accommodate. This led to many assumptions, estimates, and arbitrary allocations (Cronbach & Gleser, 1965).

Traditionally, cost-accounting systems have not established the individual's worth as a cost object, although in the accounting field, human resource accounting (HRA) represents an attempt to do so. ... The HRA movement has run aground due to the difficulties associated with operationally defining a relatively soft concept: the value of the human worker (DeAngelo, 1982; Dittman, Juris, & Revsine, 1976, 1980). Thus, HRA research has failed to provide either an acceptable method for valuing the human asset as a balance-sheet item or for calculating or estimating the value of SD_y. As a result, the accounting systems of organizations remain ill equipped to provide cost data when worth of the human asset is involved. In fact, behavioral methods are more feasible to implement in business settings (Weekley et al., 1985) because (a) the methodology required to estimate SD_y, using either the Global Estimation or the CREPID models, is specified clearly, (b) these procedures can be applied without regard to the nature of the business (profit or nonprofit; service, merchandising, or manufacturing; etc.), and (c) these procedures can be applied without regard to the type of accounting or management information system used by the firm (e.g., standard cost system or normal cost system; computerized or manual; Reilly & Smither, 1985). (pp. 588-589)

2. Global Estimation Procedure

The difficulties encountered in applying cost accounting to the estimation of SD_y eventually lead to the development of an entirely different and greatly simplified approach of estimating the standard deviation of performance. This new procedure, the global method of obtaining rational estimates of SD_y, was developed by Hunter and Schmidt (1982) and Schmidt et al. (1979) and awakened a renewed interest in utility analysis. The procedure is global in the sense of obtaining overall estimates for all employees performing at the 15th, 50th and 85th percentiles, in contrast to identifying and weighing separate components or activities underlying total dollar-valued performance for each employee separately.

Hunter and Schmidt (1982), who describe a method of obtaining SD_y estimates from experienced supervisors of budget analysts, and Schmidt et al. (1979), employ the same global procedure for computer programmers. Supervisors are used to judge productivity since they are well positioned to observe actual performance and output of employees. In the budget analyst study, supervisors were asked to estimate the yearly dollar value to their organization of the products and services provided by an average employee and also the dollar value produced by a "superior performer" at the 85th percentile. As an aid in placing a dollar value on these performance levels, supervisors

were asked to consider the cost of having an outside consulting firm provide these products and services.

Hunter and Schmidt reason that if job performance in dollar terms is normally distributed, then the difference between the value to the organization of the products and services produced by an average employee and those produced by an employee at the 85th percentile in performance is equal to SD_y. Supervisors estimated both these values and the final estimate was the average difference across the supervisors' estimates. Although the estimation task presented to the supervisors may at first appear difficult, all but one of the 62 supervisors used in the study believed that they were able to make meaningful estimates. Use of a carefully developed questionnaire to obtain the estimates may have aided supervisors significantly.

The final estimates of SD_y obtained for the budget analyst job is $11,327 per year, with a standard error of the mean of $1,120. The SD_y of $11,327 is about 60% of the average yearly salary for budget analysts at this level. The method assumes that dollar values are normally distributed. A subsequent study on estimating computer programmer dollar outcome evaluated this assumption and is described below.

The Schmidt et al. (1979) study employed 105 supervisors of federal government computer programmers. Supervisors estimated values for average programmers as well as for programmers at the 85th and 15th percentiles, providing two estimates of SD_y. The instructions given to the supervisors are shown below:

> The dollar utility estimates we are asking you to make are critical in estimating the relative dollar value to the government of different selection methods. In answering these questions, you will have to make some very *difficult judgments* . We realize they are difficult and that they are judgments or estimates. You will have to ponder for some time before giving each estimate, and there is probably no way you can be absolutely certain your estimate is accurate when you do reach a decision. But keep in mind three things:
>
> (1) The alternative to estimates of this kind is application of cost accounting procedures to the evaluation of job performance. Such applications are usually prohibitively expensive. And in the end, they produce only imperfect estimates, like this estimation procedure.
>
> (2) Your estimates will be averaged in with those of other supervisors of computer programmers. Thus errors produced by too high and too low estimates will tend to be averaged out, providing more accurate final estimates.

(3) The decisions that must be made about selection methods do not require that all estimates be accurate down to the last dollar. Substantially accurate estimates will lead to the same decisions as perfectly accurate estimates.

Based on your experience with agency programmers, we would like for you to estimate the yearly value to your agency of the products and services produced by the average GS 9-11 computer programmer. Consider the quality and quantity of output typical of the *average programmer* and the value of this output. In placing an overall dollar value on this output, it may help to consider what the cost would be of having an outside firm provide these products and services.

Based on my experience, I estimate the value to my agency of the average GS 9-11 computer programmer at _____ dollars per year.

We would now like for you to consider the "*superior*" *programmer*. Let us define a superior performer as a programmer who is at the 85th percentile. That is, his or her performance is better than that of 85% of his or her fellow GS 9-11 programmers, and only 15% turn in better performances. Consider the quality and quantity of the output typical of the superior programmer. Then estimate the value of these products and services. In placing an overall dollar value on this output, it may again help to consider what the cost would be of having an outside firm provide these products and services.

Based on my experience, I estimate the value to my agency of a superior GS 9-11 computer programmer to be _____ dollars per year.

Finally, we would like you to consider the "*low-performing*" *computer programmer*. Let us define a low-performing programmer as one who is at the 15th percentile. That is, 85% of all GS 9-11 computer programmers turn in performances better than the low-performing programmer, and only 15% turn in worse performances. Consider the quality and quantity of the output typical of the low-performing programmer. Then estimate the value of these products and services. In placing an overall dollar value on this output, it may again help to consider what the cost would be of having an outside firm provide these products and services.

Based on my experience, I estimate the value to my agency of the low-performing GS 9-11 computer programmer at _____ dollars per year. (p. 621)

The average estimated difference between the 15th and 50th percentiles was $9,955 ($SE$ = $1,035); the difference between the 50th and 85th percentiles was $10,871 ($SE$ = $1,673). The difference of $916 is about 8% of each of the estimates and is not statistically significant. The similar values for the two estimates support the assumption that the distribution was at least about normal. The SD_y values are approximately 55% of the average yearly salary for government computer programmers at this level.

Hunter and Schmidt (1982) note that their rational global estimation method has a number of desirable features: the estimation task presented to supervisors is not difficult to accomplish; empirical data show that estimates of the standard error of the mean are relatively small and that the assumption of normality of job performance is strongly supported; since estimates apply to incumbents rather than applicants, they are underestimates and thus are probably more desirable for conservative decisionmaking; averaging estimates across a large number of judges controls for idiosyncratic tendencies, biases, and random errors of individual experts; basing costs on the use of outside consultants provides a relatively concrete mental standard; and similar estimation techniques have been successfully used in high-level policy decisionmaking in other contexts.

Boudreau (1984) points out that often high precision in SD_y estimates for utility analyses may not be needed for many types of decision in the human resources context, but that only break-even SD_y values may be sufficient. Further, the results of utility analyses indicate that break-even SD_y values often fall far below global estimation values.

Much of the research accomplished to date has been directed at evaluating psychometric characteristics, e.g., reliability and validity, of global estimates. This is understandable in that the global estimation procedure is among the first methods to be developed, is most frequently used in utility analyses, and often generates very large SD_y estimates. Consequently, overall utility payoffs are very high, especially for selection programs involving many employees. Arguments have centered around six issues: accuracy of estimates; variability of estimates; difficulty on the part of supervisors in making ratings or reluctance to provide dollar-valued estimates; ambiguity in the judgment process used or confusion concerning dimensions being estimated; the normality assumption; and the lack of face validity. These issues are discussed below.

A basic consideration in evaluating a global SD_y estimate is its degree of closeness to the true or actual value of SD_y--its accuracy or validity. One measure of external validity is obtained by the direct comparison of the global SD_y estimate with an estimate using a cost accounting-based method, a measure considered the standard of comparison. Another means of measuring external validity is comparing SD_y estimates with actual standard deviations of objective performance measures such as sales (Bobko et al., 1983; Burke & Frederick, 1984; Reilly & Smither, 1985). These data considered together provide somewhat supportive but limited evidence for the accuracy of global estimates of SD_y.

A larger number of studies judge the accuracy of estimates by comparing the degree of similarity of results using alternative methods of estimating SD_y, but without reference to external criteria (Bobko et al., 1983; Burke & Frederick, 1984; Eaton et al., 1985; Reilly & Smither, 1985; Weekley et al., 1985). The results of these comparisons are mixed: some authors conclude alternatives provide disparate results; others conclude that estimates are statistically and practically close and appropriate for use in utility analyses.

With regard to the second issue, rater validity, Dreher and Sackett (1983) write that agreement among supervisors in making global estimates would at least suggest that these estimates are potentially meaningful. Yet in Schmidt's et al. (1979) pathbreaking study, very large variability in estimates was found. For example, the authors report a standard error of $1,673 for the difference between estimates of value of computer programmers between the 50th and 85th percentiles. The standard deviation of the SD_y estimate itself is $17,064--much larger than the mean of the SD_y estimate of $10,871. Bobko et al. (1983) also found substantial inter-rater variation within a set of estimates for a given percentile. Thirteen supervisors of insurance counselors provided SD_y estimates of $54,600. The standard deviation of the estimates was $48,800--a value of the same magnitude as the point estimate of SD_y. Substantial inter-rater variability or inconsistency is obviously undesirable if it reflects rater bias, sampling error, or misunderstanding of scale properties or scale dimensions.

A number of investigators suggest techniques for improving agreement among supervisors or raters. One common technique is to employ an anchor for the 50th percentile (Bobko et al., 1983; Burke & Frederick, 1984, 1986; Eaton, Wing & Mitchell, 1985). Another technique that reduces variability is to instruct raters as a group to reach consensus on point estimates or provide feedback of estimates to individuals (Burke & Frederick, 1984, 1986).

With regard to the third issue, difficulties in obtaining estimates, Hunter and Schmidt (1982) report few problems in obtaining dollar-valued global estimates from supervisors, but other investigators had different experiences. Bobko et al. (1983) state many supervisors complained that the task of estimating sales data was unduly difficult and that their estimates would be error-ridden. Eaton et al. (1985) note that 12% of tank commander supervisors refused to provide dollar estimates, stating that soldiers' lives and combat activities were not describable in dollar terms. Eaton et al. suggest dollar-valued estimating problems might also arise for civil service jobs without private industry

counterparts. Burke and Frederick (1984), Mathieu and Leonard (1987), Reilly and Smither (1985), and Rich and Boudreau (1986) also report rater difficulties.

Considering confusion in the rating process, the fourth issue, several studies indicate that the cognitive cues and process used in arriving at dollar estimates vary among supervisors. Bobko et al. (1983) note that in estimating overall worth, supervisors all mentioned salary as one appropriate indicator and some supervisors indicated that their estimates were based entirely on salary. However, Burke and Frederick (1984) state that in considering a sales manager job, salary was not the central factor. The most important dimensions in determining estimates were: management of recruiting, training, and motivating personnel; amount of dollar sales achieved based on a manager's operating budget; management of sales coverage (area of responsibility); administration of the performance appraisal system; and forecasting and analyzing sales trends. Apparently, the global estimate procedure does not provide an understanding of the process used in considering such factors as task characteristics, performance attributes, scale properties, anchors, and their interactions.

With regard to normality of distribution, the fifth issue, the global estimate procedure is based on the assumption that dollar-valued performance is normally distributed and that supervisors can estimate the difference between the value of products and services produced by an average employee at the 50th percentile and those produced by an employee at the 85th and 15th percentiles. The difference in estimates between the 85th and 50th percentiles and between the 50th and 15th percentiles provide two direct estimates of the standard deviation in dollar terms, SD_y. Schmidt et al. (1979), as previously noted, provide empirical evidence of the similarity of the two estimates that support the assumption of normality.

Bobko et al. (1983) also provide strong empirical support of underlying normality for total volume of sales used as an objective index of job performance. Other empirical studies have supported the assumption of normal distributions including Burke and Frederick (1986), Greer and Cascio (1987), and Weekley et al. (1985). Both Bobko et al. (1983) and Burke and Frederick (1984) found non-normal distributions when examining differences that included comparisons between the 85th and 97th percentiles and other point distributions. Schmidt, Mack and Hunter (1984) also found non-normal distributions in comparisons of differences at the 15th and 85th percentiles. Eaton et al. (1985) and Rich and Boudreau (1987) found percentile estimates to be non-normal.

As noted in an earlier chapter, Schmidt and Hunter (1982) argue that when performance distributions depart from normality, such departures do not seriously distort utility estimates. Bobko et al. (1983) concluded on the basis of examining normality assumptions and accuracy of SD_y that "estimates of SD_y is not necessarily the Achilles' heel of utility analyses" (p. 174).

A final issue, the sixth, is that global estimates lack face validity (Cascio, 1987a). Since the components of each supervisor's estimate are unknown and unverifiable, the procedure does not appear to measure what it purports to measure. Consequently, there is concern that global estimates of SD_y may, in practice, lack credibility among decision-makers. Eaton et al. (1985) also believe the technique lacks face validity in the military context and for some civil service jobs.

3. Estimates Based on Individual Job Performance

As previously discussed, one behaviorally based method of obtaining SD_y global estimates relies on supervisors' judgment of overall dollar value of employee performance at various points on an assumed distribution of performance. An alternate approach, the Cascio-Ramos estimate of performance in dollars (CREPID), returns to the older tradition in psychology of measuring job performance directly using carefully constructed behavioral rating scales based on job analysis results (Cascio & Ramos, 1986).

CREPID estimates the value of the work behavior of each individual as rated by supervisors on performance rating scales; the standard deviation of these values is the SD_y. The basic steps for arriving at SD_y include detailing a job in terms of its principal activities, assigning a percentage of salary to each principal activity, and then obtaining ratings on each employee's job performance or each principal activity. The ratings, in turn, are converted into estimates of dollar value for each principal activity and summed. This sum represents the value of each employee's job performance.

Cascio (1987a) points out that the method is based on the assumption that if an organization's compensation program reflects current market rates for jobs, then the economic value of each employee is reflected best in his or her salary. The method also assumes that all significant aspects of performance can be incorporated and detailed in the rating scales for principal activities and that the rating scales can properly reflect individual differences in performance.

Cascio's (1987a) summary of the procedure is shown below:

1. Identify principal activities.
2. Rate each principal activity in terms of time/frequency and importance.
3. Multiply the numerical ratings for time/frequency and importance for each principal activity.
4. Assign dollar values to each principal activity. Take an average rate of pay of participants in the study and allocate it across principal activities according to the results obtained in step 3.
5. Rate each principal activity on a 0-200 point scale.
6. Multiply (for each principal activity) its dollar value by point rating assigned (expressed as a decimal number).
7. Compute overall economic value of job performance by adding together results of step 6.
8. Over all employees in the study, compute the mean and standard deviation of dollar-valued job performance. (p. 189)

The procedure requires only two sets of ratings from a supervisor: a rating of time/frequency and importance of each activity to arrive at relative weights during the "job analysis" phase; and a rating of a given subordinate's performance on each principal activity during the "performance appraisal" phase. All other steps may be done by personnel specialists.

Cascio and Ramos (1986) applied CREPID for the positions of account supervisor in a Bell operating company. Eight principal activities defined the job including supervisory functions, receiving questions on billing problems from suppliers, adjudicating bills and vouchers, and ensuring implementation of proper security precautions. The SD_y was found to be over $10,000, about three and a half times larger than the standard deviation of the actual distribution of salaries. Cascio (1987a) notes that such high variability suggests that supervisors recognize significant differences in performance throughout the rating process.

Janz and Dunnette (1977) developed an individual, behaviorally based estimating procedure similar to CREPID. Their procedure, however, is based on estimates of the relative dollar costs associated with different levels of effectiveness on each job performance dimension, rather than estimates based on proportionally allocating salaries according to activity importance weights. A number of other studies involved direct

estimates of the value of each individual's performance without using job analysis or behaviorally based ratings, including Bobko, et al. (1983), Burke and Frederick (1984), and Simonet (1984).

Unlike global estimation, CREPID makes no assumptions regarding the underlying normality of the performance distribution and also clearly identifies the components, weights, and estimates used for each employee. The estimated value for each employee on each job activity can be evaluated for appropriateness. The process as a whole is systematic, explicit, and understandable and consequently may be credible and acceptable to users. Greer and Cascio (1987) stated that the CREPID model had more face validity than global estimates or cost accounting with a firm's top managers and accountants familiar with each procedure and the data used to make estimates. The authors conclude that the credibility of CREPID may not be a significant issue among decisionmakers, even when estimates have not been validated against some meaningful external criteria such as cost accounting outcomes.

CREPID assumes an equivalence between economic value and salary while the global estimate technique is based on the economic value of goods and services as sold. Since the economic value of goods and services is about twice the average salary (Schmidt & Hunter, 1983), the approaches, not surprisingly, yield different estimates of SD_y (Weekley, 1985; Greer & Cascio, 1987). Boudreau (1983), on the other hand, considers the appropriate economic value for utility analysis to be net benefits, the difference between sales value (revenue) and service costs (salary and benefits).

Several authors point out limitations in the mathematical characteristics of CREPID scales that may undesirably constrain estimates of SD_y (Dreher & Sackett, 1983; Reilly & Smither, 1985; Simonet, 1984). For example, CREPID employs an essentially arbitrary zero-to-200 point performance rating scale rather than a 100-to-200 point scale to reflect a performance ratio of 2 to 1, as desired, between the most and least effective employees. Simonet writes that CREPID is inadvertently scaled to produce an estimate of SD_y that is almost 70% of salary, if used correctly by raters, since the rating scale value assigned for average performance is 100 and the scale value for performance at the 85th percentile is arbitrarily set at 170. Reilly and Smither consider the scaling problem to represent a serious deficiency in previous SD_y estimates using CREPID, but that changing the scale and rater instructions can easily make the procedure conform to magnitude estimations, the original intent of the designers.

4. Estimates Based on Proportional Rules

Hunter and Schmidt (1982) were concerned that difficulties in estimating SD_y had long discouraged needed utility analyses. As one practical response to that problem, they developed a global estimation technique. Data obtained through global estimates and productivity output studies led them to recommend the use of 40% of average salary or alternatively 20% of mean output as lower bound estimates of SD_y when more detailed efforts of estimating are not feasible. Since Hunter and Schmidt's original proposal of using proportional rules of thumb for estimating SD_y, cumulative research findings strongly support their use as accurate or conservative. Since proportional rules are also simple, direct and can be used in most situations, they appear to have potential for widespread application. If 40% of salary is substituted for SD_y, utility is given in dollar terms; if 20% is substituted for SD_y, utility is given in terms of the percentage increase in output. The research evidence for proportional rules is outlined in the following section.

Hunter and Schmidt (1982) reviewed two of their other studies that used global estimates of SD_y and found the estimates to be 60% of annual salary for budget analysts and 55% for government computer programmers. They also reviewed empirical studies that gave estimates of SD_y or provided data for calculating SD_y for six different jobs (Doppelt & Bennett, 1953; Lee & Booth, 1974; Roche, 1961; Schmidt & Hoffman, 1973). The average value of SD_y as a percentage of salary for the six jobs was only 16%. However, since partial performance dimensions, e.g., tenure or training costs, were used, Hunter and Schmidt note that values of SD_y based on all facets of job performance would certainly be higher as a percentage of salary. Based on their own two studies and the six other jobs reviewed, Hunter and Schmidt estimate the true average SD_y to be in the range of 40% to 70% of salary.

Hunter and Schmidt (1982) and Schmidt and Hunter (1983) also reported that mean employee wages and salary in the U.S. economy as a whole is approximately 57% of mean output of goods and services produced. A further review of studies in the literature showed the SD_y as a percentage of salary ranged from 42% to 60%. Given these facts, the authors reasoned that the lower and upper bound predictors for the values of the standard deviation of output (SD_y) is 42% times 57% = 23.8%; and 60% times 57% = 34%, respectively.

The purpose of Schmidt and Hunter's (1983) study was to compare the predicted upper and lower bound standard deviation percentage estimates with empirical values. The data for their analysis came from studies that reported performance ratios for non-

piecework compensation systems (Lawshe, 1948; Rothe, 1946, 1947, 1970; Rothe & Nye, 1958, 1961; Stead & Shartle, 1940; Tiffin, 1947), piecerate compensation systems (Rothe, 1951, 1978; Rothe & Nye, 1959), and uncertain compensation systems (Evans, 1940; Hull, 1928; Lawshe, 1948; McCormick & Tiffin, 1974; Stead & Shartle, 1940; Wechsler, 1952).

The authors also analyzed studies that reported the mean and standard deviation of actual employee production or output rather than performance ratios for non-piecework compensation systems (Barnes, 1958; Klemmer & Lockhead, 1962), piecerate compensation systems (Barnes, 1937, 1958; Viteles, 1932) and uncertain compensation systems (Wechsler, 1952).

The results for the non-incentive condition are shown in Table 3.1; results for the piecerate system are shown in Table 3.2; and results for which the compensation system could not be determined with certainty are shown in Table 3.3.

The last column in the three tables shows the standard deviation as a percentage of mean output. The standard deviations of the compensation systems range from 0.044 to 0.052. These values are quite small; furthermore, they have not been corrected for the effects of sampling error and thus overestimate real variability across these jobs. Although Table 3.3 shows the results of studies in which the compensation system could not be determined with certainty, the authors state that the findings support their hypothesis that studies in the uncertain compensation category are based predominantly or wholly on data from jobs without piecerate compensation. Consequently, data from Tables 3.1 and 3.3 were combined.

The combined non-incentive compensation and uncertain compensation conditions yield a mean SD_y value of 20.0% of mean output (SD = 0.062). The mean ratio of 95th to 5th percentile is 2.06 (SD = 0.53). These values are 83% and 90% as large, respectively, as the predicated lower bound estimate, although the differences between each figure and lower bound estimate is statistically significant ($p < .01$).

Table 3.1. Productivity Ratios Under Non-piecework Compensation Systems

Study	Job	*N*	Ratio of 95th to 5th percentiles	Ratio of *SD* to average
Klemmer & Lockhead (1962) [a]				
Study 1	Card punch operators	Not reported	1.47	0.116
Study 2	Proof machine operators	Not reported	1.57	0.135
Rothe (1946)	Dairy workers	8	2.23	0.232
Rothe (1947)				
Time 1	Machine operators	130	1.90	0.189
Time 2	Machine operators	130	2.69	0.278
Time 3	Machine operators	130	2.80	0.288
Rothe & Nye (1958)	Industrial workers	27	1.94	0.194
Rothe & Nye (1961) [b]				
1958	Machine operators	37	1.77	0.169
1960	Machine operators	61	1.41	0.103
Rothe (1970)	Welders	25	1.91	0.190
Stead & Shartle (1940)	Typists	616	1.89	0.187
Lawshe (1948)	Cashiers	29	1.95	0.196
Tiffin (1947)	Electrical workers	33	1.83	0.178
Barnes (1958) [a]	Assembly workers	294	1.60	0.140
M			1.93	0.185
SD			0.40	0.052

Source: Schmidt and Hunter (1983), p. 409

[a] Means and standard deviations for experienced workers given; ratios are based on these.

[b] *N*s are averages across 11-12 weeks.

Table 3.2. Productivity Ratios Under Piecerate Compensation Systems

Study	Job	*N*	Ratio of 95th to 5th percentiles	Ratio of *SD* to average
Rothe (1951)	Candy manufacture	18	1.50	0.122
Rothe & Nye (1959)	Machine operators	42	2.35	0.245
Rothe (1978) [a]				
Department A	Foundry workers	26	1.66	0.151
Department B	Foundry workers	19	1.43	0.108
Department C	Foundry workers	17	1.52	0.125
Department D	Foundry workers	21	1.45	0.112
Barnes (1958) [b]	Assembly workers	314	1.42	0.105
Barnes (1937) [b]	Lathe operators	121	1.63	0.146
Tiffin (1947)				
Group 1	Electrical workers	39	2.08	0.213
Group 2	Hosiery loopers	99	1.91	0.190
Viteles (1932) [b]	Weavers	239	1.58	0.137
M			1.69	0.150
SD			0.29	0.044

Source: Schmidt and Hunter (1983), p. 410

[a] Mean *N*s: *N*s varied slightly over weeks.

[b] Means and standard deviations given; ratios are based on these.

Table 3.3. Productivity Ratios in Studies with Uncertain Compensation Systems

Study	Job	*N*	Ratio of 95th to 5th percentiles	Ratio of *SD* to average
Hull (1928 p.35)				
Group 1	Shoe factory workers	NR	NC	
Group 2	Hosiery factory workers	NR	NC	
Group 3	Textile industry workers	NR	NC	
Group 4	Shoe factory workers	NR	NC	
Group 5	Textile industry workers	NR	NC	
Group 6	Silverware manufacture	NR	NC	
Evans (1940) [a]	"A Number of handcraft jobs"	NR	2.22	0.230
Stead & Shartle, (1940)				
Group 1	Sales clerks	153	3.46	0.335
Group 2	Card punch operators			
	Day shift	113	1.62	0.144
	Night shift	121	1.80	0.174
Group 3	Lamp shade manufacturer	19	1.47	0.116
Group 4	Card punch operators	62	2.84	0.291
Group 5	Sales clerks	NR	NC	
Group 6	Sales clerks	NR	NC	
Lawshe (1948)				
Group 1	Drilling	11	3.37	0.330
Group 2	Wool pullers	13	2.00	0.203
Group 3	Sales clerks	18	17.35	0.542
Wechsler (1952) [b]				
Group 1	Machine workers	101	1.96	0.197
Group 2	Electrical workers	100	1.54	0.129
Group 3	Electrical workers	65	1.78	0.171
McCormick & Tiffin (1974)				
Group 1	Cable workers	40	2.29	0.238
Group 2	Electrical workers	138	1.91	0.190
Group 3	Assemblers	35	2.54	0.264
M [c]			2.20 (3.21)	0.215 (0.237)
SD [c]			0.62 (3.83)	0.067 (0.104)

Source: Schmidt and Hunter (1983), p. 411

Note: NR = not reported; NC = not computable

[a] Adjusted ratio computed based on information that for all of these jobs the standard deviation of output was approximately 23% of mean output

[b] Means and standard deviations given; ratios are based on these.

[c] Values in parentheses include the Lawshe (1948) sales clerks; values not in parentheses do not.

Schmidt and Hunter note that the findings primarily reflect quantity of output; quality is probably reflected only crudely. If quality of output were properly incorporated, the *SD* of output would be greater, since quality and quantity are positively correlated. The findings reported are based on blue collar skilled and semiskilled jobs and lower level white collar jobs. The authors' earlier estimates of 23.8% to 34% were based in large part on middle level jobs allowing for very expensive errors, in turn increasing the standard deviation disproportionately relative to salary (Hunter and Schmidt, 1982). These findings lead the authors to conclude that for jobs without incentive based compensation:

> the standard deviation of employee output [can be estimated] at 20% of mean output without fear of overstatement. This figure is probably somewhat conservative. . . . The findings of this study provide support for the practice that we have recommended of estimating SD_y (the standard deviation of output in dollars) as 40% of mean salary . . .. (p. 412)

Hunter, Schmidt, and Judiesch (1989) extended the Schmidt and Hunter (1983) by examining the variability of employee output as a percentage of mean output (SD_p) as a function of the complexity level of the job. Other refinements from the previous study were adjustments of observed SD_p figures for the inflationary effects of measurement error and the deflationary effects of range restriction. They found that SD_p increases as the information processing demands (complexity) of the job increases. Progressing from low to medium to high complexity non-sales jobs, SD_p is 19%, 32%, and 48%, respectively. SD_p values were 120% for life insurance sales jobs, and 52% for non-insurance sales jobs.

In addition to the reviews of Hunter and Schmidt (1983) and Schmidt and Hunter (1982), a number of recent studies have reported SD_y expressed as a percentage of salary. Most of these studies show SD_y in the 40% to 70% range of subjective estimates of worth. (Bobko et al., 1983; Cascio & Silbey, 1979; Greer & Cascio, 1987; Reilly & Smither, 1985; Rich & Boudreau, 1986; Schmidt, Mack & Hunter, 1984; Weekley et al., 1985).

A few studies report SD_y estimates above 70% in the military setting (Eaton, Wing & Mitchell, 1985) and in comparison with sales or revenues (Burke & Frederick, 1984; Reilly & Smither, 1985).

Also a few studies that base SD_y estimates on salary or a partial criterion of value produce values lower than 40% (Cascio & Ramos, 1984; Hunter & Schmidt, 1982). In only two studies, employing a criterion reflecting full value, were SD_y estimates under 40% found (Arnold, Rauschenberger, Soubel & Guion, 1982; Mathieu & Leonard, 1987).

Thus, based on comparisons involving more than two dozen jobs, Hunter and Schmidt's 40% to 70% mean salary proportional rule can be safely considered as accurate or conservative.

For the mean output proportional rule, 23% to 34%, a review of recent studies, using subjective estimates of output, shows all SD_y estimates above 23% (Bobko et al., 1983; Burke & Frederick, 1984; Cascio & Ramos, 1986; Eaton et al., 1985; Rich & Boudreau, 1986; Schmidt, Mack & Hunter, 1984; Weekley et al., 1985). Again, Hunter and Schmidt's lower bound 20% output proportional rule can be safely considered as accurate or conservative.

The data taken as a whole indicate that these proportional rules, applied uniformly to selection alternatives, may safely be used as conservative estimations; however, in very large selection and classification programs such as in the military, estimates produced by proportional rules would likely result in very low estimates of overall dollar value.

As Boudreau (1984) and Burke and Frederick (1986) note, decisions regarding the adoption of one versus another personnel selection procedure are not likely to be modified by the SD_y estimate used since they are based on comparisons of alternative selection procedures. When comparisons are made among various types of interventions, (e.g., selection, training, design changes) decisions are more likely to be affected by the type of SD_y estimate used in determining utility.

5. Superior Equivalents Techniques

Eaton, Wing and Mitchell (1985) developed two alternative techniques, "superior equivalents" and "systems effectiveness," for estimating dollar value of performance that appeared to be more useful in certain job contexts, particularly in the military. Eaton et al. were concerned that the more conventional global method might be used in situations where estimation would be "impractical, if not misleading . . . where the nature of the work is such that managers are more accustomed to consider the relative productivity of employees or crews than the costs of producing given levels of output. Such situations could also occur where employees operate very complex, expensive equipment and/or are focal to the productivity of a costly system" (p. 29).

The two techniques share the CREPID characteristic of evaluating performance, but without a job analysis; they share the global estimation characteristic of comparing overall

performance of individuals at the 85th and 50th percentiles, but use an anchor for mean performance.

The superior equivalents technique uses supervisor estimates of the number of superior (85th percentile) performers that would be needed to produce the output of a fixed number of average (50th percentile) performers. However, in place of using direct estimates of the dollar value of 85th percentile performers as in global estimates, Eaton et al. obtained estimates of the number of superior tank commanders needed to equal the performance of a standard company of 17 tanks with average performance. The dollar value of average performance is based on either actual compensation (salary and benefits) or estimates made by tank commanders and expressed in equivalent civilian salary.

Underlying this technique, then, is the belief that supervisors can make more accurate judgments of relative performance than direct estimates of the dollar value of that performance.

Where the dollar value of average performance (V 50) is known, or can be estimated, the standard deviation in dollars, SD_y, may be estimated by using the ratio N 50/N 85 times V 50 to obtain V 85, and then subtracting V 50. This can be shown as:

$$SD_y = V\,85 - V\,50$$

and,

$$V\,85 = (V\,50 \times N\,50) \div N\,85$$

Hence,

$$SD_y = V\,50[(N\,50 \div N\,85) - 1] \quad . \tag{3.1}$$

A questionnaire was developed by Eaton et al. to obtain estimates of the number of tanks with superior tank commanders needed to equal the performance of a standard company of 17 tanks with average commanders, as well as estimates of dollar value. For both number-of-tank and dollar value, a fill-in-the-blanks format was used. The portion of the questionnaire concerning tanks is shown below:

> For the purpose of this questionnaire an "average" tank commander is an NCO or commissioned officer whose performance is better than about half his fellow TCs. A "superior" tank commander is one whose performance is better than 85% of his fellow tank commanders.
>
> The first question deals with relative value. For example, if a "superior" clerk types 10 letters a day and an "average" clerk types 5 letters a day then,

all else being equal, 5 "superior" clerks have the same value in an office as 10 "average" clerks.

In the same way we want to know your estimate or opinion of the relative value of "average" vs. "superior" tank commanders in combat.

1. I estimate that, all else being equal, _____ tanks with "superior" tank commanders would be about equal in combat to 17 tanks with "average" tank commanders. (p. 33)

Computing average equivalent salary at $30,000, collecting questionnaire data on 100 advanced training tank commanders and employing Equation (3.1), SD_y was found to be $26,666. The authors believe that the technique worked well and provided consistent and accurate estimates of the number of superior performers needed to equal the aggregate performance of a fixed number of average performers.

Two concerns were expressed by the authors concerning both techniques. The techniques are based on the assumption that performance is largely a function of the performance of the commander and also that performance quality in some situations may not be easily linked to a unidimensional, quantitative scale.

6. Systems Effectiveness Technique

A second alternative method developed by Eaton, Wing and Mitchell (1985) was designed for use in work situations where the cost of equipment operated by the employee was considerably greater than the individual's salary, e.g., an army tank commander. The technique considers aggregate performance as a function of the number of employee/machine units and the quality of performance of the units. The value of improved aggregate system performance is defined as being equal to the cost of the increased number of units needed to obtain that aggregate level of performance.

Eaton et al. modified Brogden's utility formulation so that SD_y could be more readily estimated by cost and performance measures. The authors state that frequency-type variables, e.g., hits per firing (army tank commander), number of convictions per year (detective), and number of pupils achieving a given standard (teacher) may be useable in this formulation. The assumption, as made earlier, is that the performance of the unit in the system is largely dependent on the performance of the employees in the job under study.

As in the superior equivalents technique, the authors differentiated between the standard deviation in dollar terms, $SD\$$, and the standard deviation in output units of

performance, SD_y. They devised a dollar estimation based on Brogden's formulation that showed:

$$SD\$ = (CuSD_y)/Y\,1 \qquad (3.2)$$

where Cu is the cost of the unit in the system (tank purchase costs, maintenance, and personnel); SD_y is the standard deviation of output units such as hits per firing from a tank commander; and $Y\,1$ is the level of performance of average performers.

The payoff scale, expressed as cost savings, is similar to the proportional rule of computing the SD in dollars based on percentage of average salary, but in the Eaton et al. formulation including salary as one element in the total cost of a unit.

Results of previous criterion-related validity studies on tank crew performance showed validities (used in the current study to estimate mean performance levels) ranging from 0.2 to 0.5, and accounting data showed total tank costs ranging from $300,000 to $500,000 per year. Applying conservative values of $300,000 for costs, 0.2 for validity and 0.2 for the ratio of $SD_y/Y1$ in Equation (3.2), $SD\$$ was computed to be $60,000. Both techniques were found to produce larger estimates, but smaller variability than global estimates.

C. EMPIRICAL COMPARISONS OF ALTERNATIVE, SD_y ESTIMATES

If there were appropriate and available criteria against which to validate SD_y estimates, the need for using behaviorally based methods, in the first place, would be removed. Comparing one behaviorally based estimate with another, especially in the absence of an actual real-world selection decision context, is unlikely to provide definitive support of one method over others (Weekley et al., 1983). This section reviews a number of empirical studies comparing various methods of estimating SD_y to obtain a sense of relative adequacy pertaining to consistency and agreement. Several studies employ objective measures and permit more than tentative statements of accuracy.

1. Estimation of SD_y Employing An Objective Criterion

In an empirical check on the ability of judges to generate normally disturbed global estimates and to evaluate the accuracy of these estimates, Bobko et al. (1983) used two measures of performance: an objective, specific estimate of actual yearly dollar sales obtained from archival records (number of policies sold times average policy value); and supervisors' estimates of overall performance that reflected the type of measure used by

Schmidt et al. (1979). Supervisors' estimates of specific yearly dollar sales were also obtained and the three measures compared.

Estimates at the 15th, 50th, 85th, and 97th percentiles of overall worth and specific dollar sales were obtained from 17 supervisors of 92 insurance counselors in a large insurance company.

Table 3.4 gives the estimate for yearly sales based on supervisors' ratings and archival data. Three estimates of SD_y can be directly estimated from the table. For the actual archival sales volume data, the mean was $124,882, the median was $117,300, and the overall standard deviation was $52,308. The distribution of actual sales data did not depart from normality. The standard deviation estimated from the 85th and 97th percentiles ($30,000) was significantly different from $52,308. However, the two estimates within the 15th percentile to 85th percentile range were not statistically different.

Table 3.4. Estimated Percentiles and Standard Deviations for Yearly Sales, in Thousands of Dollars

Supervisor	15%	SD_y (50%-15%)	50%	SD_y (85%-50%)	85%	SD_y (97%-85%)	97%
1	35.0	70.0	105.0	35.0	140.0	17.5	157.5
2	45.0	30.0	75.0	45.0	120.0	30.0	150.0
3	26.0	59.0	85.0	69.0	154.0	54.0	208.0
4	78.0	130.0	208.0	182.0	390.0	78.0	468.0
5	48.7	97.4	146.1	48.7	194.8	97.4	292.0
6	50.0	50.0	100.0	100.0	200.0	25.0	225.0
7	37.5	37.5	75.0	25.0	100.0	25.0	125.0
8	75.0	225.0	300.0	98.0	398.0	3.0	401.0
9	53.0	35.0	88.0	52.0	140.0	35.0	175.0
10	4.0	12.0	16.0	8.0	24.0	2.0	26.0
11	4.8	14.4	19.2	22.8	42.0	3.2	45.2
12	10.6	3.3	13.9	5.9	19.8	7.7	27.5
13	10.0	7.0	17.0	19.0	36.0	12.0	48.0
M	36.7	59.3	96.0	54.6	150.7	30.0	180.6
SD	24.9	62.0	83.2	48.8	124.1	29.9	139.7
Actual	70.0	47.3	117.3	55.6	172.9	52.4	225.3

Source: Bobko, Karren, and Parkington (1983), p. 173.

Note. The overall mean (in thousands of dollars) was 124.9. The overall standard deviation was 52.3.

Table 3.5 gives the supervisors' ratings of overall worth. Again, these estimates of SD_y can be directly estimated from the table. The estimated worth of the average employee was $16,000. One pairwise difference among the three estimates of SD_y was statistically significant: the difference between the estimate computed from the 85th and 97th percentiles ($3,800) and the estimate computed from the 50th and 85th percentiles ($6,400).

Table 3.5. Estimated Percentiles and Standard Deviations for Value of Overall Products and Services, in Thousands of Dollars

Supervisor	15%	SD_y (50%-15%)	50%	SD_y (85%-50%)	85%	(97%-85%)	SD_y 97%
1	9.5	4.0	13.5	2.0	15.5	1.0	16.5
2	10.4	1.3	11.7	1.8	13.5	2.1	15.6
3	12.5	27.5	40.0	39.0	79.0	16.0	95.0
4	12.0	3.0	15.0	3.0	18.0	2.0	20.0
5	6.6	0.7	7.3	2.8	10.1	2.9	13.0
6	10.0	6.0	16.0	4.0	20.0	2.0	22.0
7	11.5	1.0	12.5	1.5	14.0	1.0	15.0
8	12.0	2.0	14.0	2.5	16.5	0.9	17.4
9	10.4	1.3	11.7	2.1	13.8	2.1	15.9
10	13.4	5.2	18.6	3.9	22.5	1.0	23.5
11	11.2	3.3	14.5	3.7	18.2	1.3	19.5
12	10.6	3.4	14.0	2.5	16.5	1.5	18.0
13	16.0	2.6	18.6	15.4	34.0	15.0	49.0
M	11.2	4.7	16.0	6.4	22.4	3.8	26.2
S	2.2	7.0	7.8	10.4	18.0	5.2	22.6

Source: Bobko, Karren, and Parkington (1983), p. 173.

The authors conclude that averaging supervisors' judgments may provide adequate point estimates of actual variation in performance, at least when an objective measure of performance is used and estimates are within the 15th to the 85th percentile range. Also, the assumption of underlying normality for the objective measure of performance was strongly supported.

The authors note, as can be seen in Tables 3.4 and 3.5, substantial variations within any given set of estimates for a particular percentile or estimate of SD_y. For example, the estimate for SD_y in Table 3.4 (comparing 85th and 50th percentiles) is $54,600, but the standard deviation of the 13 supervisors' estimate is $48,800. Bobko et al. suggest the use of a sequential estimation procedure to reduce variability that would be based on Hogarth's (1981) discussion of the importance of feedback in judgmental heuristics.

The authors also note that "Many supervisors complained that the task of estimating sales data was unduly difficult and would be error-ridden" (p. 175). The estimated values obtained were far lower than the actual sales standard deviations. When the supervisors were asked how they arrived at the overall worth estimates, all mentioned salary. In all, Bobko et al. express optimism concerning the use of SD_y estimates in utility analyses.

2. Comparisons of SD_y Estimates Based on Feedback Procedures

Burke and Frederick (1986) compared estimates based on two consensus-seeking procedures, global estimation, and proportional rules. As in the earlier work of Bobko et al. (1983) and Burke and Frederick (1984), this study included the 97th percentile as a fourth point estimate. A district sales manager's job in a large national manufacturing organization was selected for analysis.

For the global estimation method, regional sales managers (one level above district sales managers) were initially asked to estimate the annual value of service provided by district sales managers at the 15th, 50th, 85th and 97th percentiles. The Schmidt et al. (1979) procedure was followed in developing instructions to obtain global estimates. The averages for the three differences were obtained, yielding three SD_y values. The final estimated value of SD_y was the average of either two or three SD_y estimates.

For procedures A (group consensus) and B (individual feedback), the average 50th percentile estimate was fed back to managers in two ways. For procedure A, four regional managers were given a computed average estimate and then instructed to reach consensus as a group. For procedure B, 22 regional sales managers were given a computed average estimate and then asked individually to provide the other three point estimates.

Estimates of SD_y based on the proportional rule of 40% and 70% of average salary were also computed.

Table 3.6 shows the SD_y estimate for each of the procedures. The authors conclude that of the eight SD_y estimates calculated, a tentative case can be made for using the SD_y estimates based on four percentile points for procedure B or for the global procedure. Values for these two procedures were virtually identical, $32,323 and $32,287, respectively. The global procedure was the most stable of the three behaviorally based procedures when going from three to four percentile points, percentage changes ranging from about 39% for procedure A to about 8% for the global procedure.

Table 3.6. Assessment Center SD_y Estimates for Various Procedures

SD_y estimation procedure	SD_y
Based on 15th, 50th, and 85th percentile points	
Procedure A	27,500
Procedure B	28,151
Schmidt, Hunter, McKenzie and Muldrow (1979)	35,192
Based on 15th, 50th, 85th, and 97th percentile points	
Procedure A	38,333
Procedure B	32,323
Schmidt, Hunter, McKenzie and Muldrow (1979)	32,287
Percentage of mean salary	
40% of mean salary	12,789
70% of mean salary	22,381

Source: Adapted from Burke and Frederick (1986), p. 337

The authors note that the shape of the true performance distribution implied by procedures A and B differed from that for the global procedure, i.e., the estimated distribution for procedures A and B were positively skewed, suggesting that the true performance distribution might be positively skewed. The authors also note that procedure A reduced within-column variations compared to the global procedure more than procedure A. But they note that decisions regarding the use of one versus another selection procedure are not likely to be affected by the SD_y estimate employed.

Burke and Frederick argue that estimates are likely to vary for the same job depending on how judges define the payoff function and that thus there is a continuing need to determine what dimensions and information raters use in making SD_y estimates.

3. Comparison of Superior Equivalents and Systems Effectiveness Technique with Conventional Estimating Methods

Eaton, Wing and Mitchell (1985), as described earlier, developed new approaches for estimating SD_y that were based on the change in the number and performance levels of system units which lead to increased aggregate performance. The focus of their method is a system comprised of tanks and tank crewmen in the Army. In this context, tank commanders (TCs) are more attuned to evaluating relative performance of soldiers or crews than to overall worth of output in dollar terms.

In the superior equivalents technique, estimates are obtained on the number of superior (85th percentile) performers needed to produce the output of a fixed number of average (50th percentile) performers. These estimates, when multiplied by the known (or estimated) dollar value of average performance, provide an estimate of SD_y. When average value is defined as average compensation, the underlying payoff scale reflects savings in personnel costs.

In the systems effectiveness technique, the value of an aggregate performance improvement due to increased performance of a fixed number of tanks is tied to the value of increased number of tanks needed to achieve comparable performance. The underlying payoff scale reflects savings in tank costs.

Eaton et al. developed a questionnaire following the procedures used by Schmidt et al. (1975) to obtain estimates of the dollar value of average and superior TC performance. The questionnaire also was used to obtain estimates of the number of tanks with superior TCs that were needed to equal the performance of a standard company of 17 tanks with average TCs. Data were obtained from two groups of TCs enrolled in advanced training with 9 to 10 years of experience as tank crew members.

To obtain an independent value of average performance, published tables of pay and allowances were used. In 1983, base pay for relevant years of experience and rank ranged from $14,000 to $26,000 and benefits could amount to more than $10,000 for typical TCs. Equivalent civilian salary was estimated at about $30,000 per year.

Previous research on tank crew performance suggested that meaningful values for the ratio $SD_y/Y1$, the standard deviation of performance (SD_y) to the initial or average level of performance ($Y1$), range from 0.2 to 0.5. The authors selected the lower bound value, a value consistent with those found in the review of output by Schmidt and Hunter (1983). Tank costs, including purchase costs, maintenance, and personnel, were estimated to range between $300,000 to $500,000. The lower bound value was also selected; both the ratio value and costs were used to compute the standard deviation in dollars, $SD\$ = (Cu\ SD_y)/Y1$, as described earlier.

Table 3.7 shows the standard deviation dollar estimates computed. The authors note that the global estimates at both the 50th and 85th percentiles for both samples of TCs suffered from considerable positive skewing and there was minimal agreement either within or between groups. These defects, the authors conclude, make the global estimates "highly

suspect" and were due in all probability to the difficulty of making dollar valued judgments in the military context.

Table 3.7. Estimates of *SD*$ for Various Techniques

	n	*SD*$ [a]
SD$ Estimation Technique		
Group 1	48	$20,000
Group 2	40	$60,000
Superior Equivalents Technique		
Using Pay and Allowance Estimates of *V* 50		
Group 1	52	$26,700
Group 2	45	$26,700
Using *SD*$ Estimates of *V*50		
Group 1	52	$26,700
Group 2	45	$31,100
System Effectiveness Technique	-	$60,000
Salary Percentage Technique	-	$12,000

Source: Adapted from Eaton, Wing and Mitchell (1985), p. 35.

[a] Rounded to nearest $100.

The median response given for the number of superior TCs equivalent to 17 average TCs was 9; the mode was 10 in both groups. The response "9" was judged to be the appropriate value and used to determine *SD*$ by the superior equivalents technique, where $SD\$ = V\,50[(N\,50/N\,85) - 1]$, as described earlier. Given average compensation of $30,000 from the pay and allowance tables, a superior TC would be valued at $30,000 times 17/9 or about $56,700. Thus the *SD*$ for the superior equivalents technique, using pay and allowance tables, is shown in Table 3.7 as $26,700. Using global estimates of average value ($30,000 and $35,000 for the two samples) yields *SD*$ of $26,700 and $31,100, respectively.

The use of average compensation for the estimate of average value can be translated into an estimate of payroll savings. When only 9 superior TCs are needed instead of 17, payroll costs are reduced from $510,000 (17 × $30,000) to $270,000 (9 × $30,000). Thus, the estimate of SD_y of $26,700 reflects the payroll savings from replacing 1.89 average TCs with one superior TC ($270,000/9 = $26,700). The authors conclude, as evidenced by the consistent and restricted number of superior performers estimated, that the

superior equivalents technique provides accurate and believable estimates of the standard deviation in dollar terms.

As shown in Table 3.7, the systems effectiveness technique produced a *SD* $ = $60,000. The authors note that while their unit cost estimates are crude, they may fairly accurately reflect reality since the estimates are based on well understood performance and cost data. Of course, acceptance of the systems effectiveness estimates is dependent on the extent one assumes that the performance of the tank is largely a function of the performance of the commander.

Eaton et al. suggest that the superior equivalents method underestimates value since raters appear to judge worth principally on the basis of salary. Perhaps this is so because the judges could not easily assign values to intangible job components and/or could not estimate the cost of contracting the work out in the civilian sector. Both the estimates of value of average performance based on the global estimates and on the tables of pay allowances, result in estimates similar to those obtained using the CREPID approach in other studies. Thus, if wages are 57% of output, as discussed earlier, then, if the *SD*$ estimate of $30,000 is multiplied by 1.75 (1/0.57), *SD*$ = $52,500--a value much closer to the one obtained by using the systems effectiveness technique.

The 40% of average salary proportional rule produced a *SD*$ = $12,000, a value much smaller than those obtained by all the other methods. The authors suggest the differences in values may be due to the greater responsibility inherent in the job of tank commander as captured by the new techniques, i.e., the incumbents can enhance the productivity of subordinates. This reasoning is similar to Schmidt and Hunter's (1983) view that the standard deviation of output should be larger for higher level jobs.

4. Comparison of Two Behaviorally Based Methods with Cost Accounting

Greer and Cascio (1987) look toward the accounting sector to develop objective, verifiable, and reliable outside criteria. The authors developed a cost accounting-based approach that they considered merely a start in the direction of contrasting behaviorally based SD_y estimates with external criteria. In their highly significant study a cost accounting estimate was used as the standard of comparison with estimates based on global estimation and CREPID.

The study was conducted in a soft drink bottling company that manufactures, merchandises, and distributes nationally known products. The job of route salesman was

chosen for analysis because the large number of individuals employed in the job and the variability in performance levels create significant differences in payoff for both the company and individual route salesmen.

The cost accounting method consisted of a carefully developed eight-step procedure that, except for one step, was supported by conventional managerial accounting theory. A nonconventional subjective procedure was used in one step to determine the degree of influence that a salesman has over the output level on his route. The authors note that integrating the subjective step with the other objective steps may diminish the cost accounting estimates as a measure of truth against which the two behaviorally based measures are evaluated.

The authors characterize the procedure used as a "profit contribution" approach to costing performance as contrasted to Roche's (1961) "profit attribution" approach, the advantage of the profit contribution approach being that no arbitrary allocation of fixed costs, essentially costs of being in business, are attributed to employees. Contribution costing generally is recommended for internal, managerial reporting purposes, but not for external reporting purposes.

Global estimates were obtained following the questionnaire-based procedure developed by Schmidt et al. (1979). CREPID estimates were obtained following the procedure developed by Cascio and Ramos (1986). Estimates were provided by 29 supervisors on 62 route salesmen. Data also were obtained from the accounting records of the firm.

Table 3.8 shows the major results obtained. The magnitude and direction of the relations among the three estimates of SD_y are apparent, even without statistical tests. The authors note that proponents of the global estimation approach are sure to be encouraged by the results of their study. One key reason is that the global estimate of SD_y is quite accurate. The global estimate was found to be \$14,636 compared to the cost accounting estimate of \$15,864 (92% of the cost accounting estimates of SD_y).

A second reason is the impressively high degree of multi-method convergence of estimates found. The global estimate of SD_y was found to be 1.6 times larger than the CREPID estimate of SD_y. Global estimates are based on the value of output as sold; CREPID estimates are based on salary. Schmidt and Hunter (1983), as previously described, calculated that in the U.S. economy, wages average 57% of the value of output.

Table 3.8. Standard Deviation Estimates Using Three Methods

Method	SD_y estimate	*M* worth	*M* value	Range	Test of significance		
					SD_{y2}	SD_{y3}	SD_y
Cost accounting (SD_y)	$15,864 (100%)	$44,985 (100%)	$41,166 (100%)	$11,237-$99,557	90% CI (*ns*)	-	-
Global estimation model (SD_{y2})	$14,636 (92%)	$31,979 (71%)	$23,000 (56%)	0-$175,000	-	90% CI (ns)	-
CREPID (SD_{y3})	$8,988 (57%)	$38,435 (85%)	$40,489 (98%)	$16,855-$53,171	-	-	t(60) = 5.344, $p < 0.10$

Source: Greer and Cascio (1987), p. 592.

Note. The percentage figures after each of the dollar values represent the dollar values expressed as a percentage of the corresponding cost-accounting-based value. SD_y = standard deviation of job performance; CI = confidence interval; CREPID = Cascio-Ramos estimate of performance in dollars; *M* worth = mean value of performance; and *M* value = median value of performance.

Thus, the CREPID estimate should be about 57% of the value of the global estimate and of the cost accounting estimate which is also based on worth. Table 3.4 shows the CREPID estimate to be exactly 57% of the cost accounting approach and 61% of the global estimation approach. Conversely, as the authors point out, estimates produced from both the cost accounting and global estimation approaches should be about 1/0.57 = 1.75 times as large as the CREPID estimate--and they are.

A third reason to be encouraged by the results obtained by the global estimation approach is that it produced a SD_y estimate that was 55% of the estimate of average salary of route salesmen. This estimate falls within the 40% to 70% range suggested by Schmidt and Hunter (1983).

Further support is given to the proportional rule for estimating SD_y in dollars as 40% of salary. This approach produces values that fall within the range of the other three estimates. The actual average wage paid route salesmen was $26,585. Applying the proportional rule, a SD_y estimate of $10,634 is produced. This estimate falls between the global estimation value of $14,636 and the CREPID value of $8,988. The proportional rule estimate is 67% of the cost accounting estimate of $15,864. Thus, Schmidt and Hunter's suggestion that the 40% of salary rule may safely be used as a conservative estimate is confirmed in this study.

The actual distribution of output in this study was positively skewed. The global estimation method, however, indicated normality of performance and thus failed to detect skew.

The global estimation approach, however, resulted in excessive variability in the point estimation of worth and also caused some confusion in raters concerning interpretations of dimensions used in estimating worth. In contrast, Greer and Cascio point out that CREPID produced a much tighter range of values than did the other two methods, and had the highest degree of face validity among those in the organization familiar with all three methodologies.

D. ESTIMATING SD_y AND DECISIONMAKING

Classical utility is evaluated in terms of the predictive efficiency implied by the validity coefficient; modern decision theoretic selection adds dollar-valued performance and the selection ratio. Value, then, depends on the interrelations among the three variables within a specific decision context.

A major obstacle in undertaking decision theoretic utility analyses appeared to have been removed about a decade ago with the development of a practical means of estimating SD_y. Many believed that an era of almost total preoccupation with the magnitude of the r_{xy} parameter was to be replaced by a new focus--utility analyses. Thus far, however, research interest has been directed primarily toward measurement of the SD_y parameter, not toward utility analyses of alternate selection procedures in a true decision context. It seems that one partial measure of value, r_{xy}, was replaced by another partial measure, SD_y.

Many of the empirical studies conducted compare various estimating techniques to one another in terms of their psychometric characteristics, including inter-rater variability, consistency with other measures, normality of the underlying performance distribution, nature of the dimensions being measured and accuracy of estimates. Few studies address the issue of how utility information is being used by decisionmakers in operational selection programs.

New techniques or modifications in estimating SD_y are advanced generally on the basis of having greater face validity or credibility, even though they are more complex and difficult to use. Given that gauging accuracy of estimates will continue to be difficult without objective, external criteria, and that decisions regarding selection procedures are not likely to be affected by the type of SD_y estimate employed, advancing one measure over another is hard to defend. It may be more worthwhile now to carry out analyses of alternative selection procedures in a problem-oriented organizational decision context so as to understand better truly significant factors in the decision and implementation process.

Comparisons among methods of estimating SD_y clearly show that different methods produce different SD_y values. Although differences in estimates may be small, statistically and practically, there is some concern that even small differences can significantly overestimate or underestimate aggregate utility. There is now a sufficient body of comparative findings to give some indications of relative trends among behaviorally based estimates. Some broad trends are briefly noted below.

Proponents of global estimation can be encouraged by comparative results. Global estimates are found to be quite accurate in the few studies that compared them with actual, objective estimates. In most other comparisons, global estimates are found to be conservative, but show high rater variability.

CREPID produces estimates that are found to be about one-half the size of global estimates, but are believed to have high credibility.

The mean proportional salary rule of 40% to 70% almost always produces very conservative results while the mean output proportional rule of 23% to 34% almost always produces estimates that fall within that range.

The superior equivalents and systems effectiveness techniques appear to produce larger estimates than global estimates, but produce smaller inter-rater variability. These two techniques may be more appropriately applied to higher level jobs that impact on the productivity of a group or system or in situations where estimates of relative performance are more natural indicators than overall worth as sold.

Reduction in inter-rater variability is made possible by anchoring the 50th percentile and by seeking consensus through feedback. Employing the 97th percentile as an additional point in averaging SD_y, estimates show mixed results. The dimension that raters rely on most heavily in making estimates of worth of employees appears to be salary, although the process of making such estimates is sometimes found to be confusing or difficult.

Differences in estimates are, of course, attributable to more than the use of different estimation methods. Differences in payoff scales, reflecting different outcomes, produce different results. Some controversy remains concerning appropriate economic outcomes to use in utility analyses. The major elements of economic performance are efficiency, technological progress and equity in distribution. The basic meaning of efficiency, the element of primary interest in the utility context, is that it yields a maximum value of outputs from any given total of inputs. Both physical quantities and economic values (prices) are involved in defining best conditions. Selection programs relate to internal efficiency attained by excellent management within the firm. The common sense of this is clear and familiar. Managers use all ways to attract, select, train, motivate and utilize employees, and to cut costs and keep operations lean.

One common approach to the improvement of internal efficiency is to improve the quality of the workplace through selection programs. Selection programs can be shown to pay off by enhancing revenues or by cost reduction. The overall aim of selection is to increase quality and quantity of products and services at minimum costs.

A number of productivity-based payoff scales are useful indices of dollar-valued performance, including output, value of goods and services as sold, sales, profits, and cost reductions. A number of partial payoff scales may also be useful such as training success or tenure. The appropriateness of the payoff scale depends on the decision context.

Often it may be necessary to examine separately all the components contributing to net productivity dollar value gain of a policy. For example, outcome values in a military selection and classification program contributing to net productivity gains may include increases in predicted performance, reduced training attrition, and increased recruiting costs. Decisionmakers may wish to reexamine the outcome value results to understand better how outcomes were weighted and combined. Based on such a reconsideration, decisionmakers may not wish to implement a policy based on reduced recruiting costs (by lowering entrance standards) resulting in reduced levels of predicted performance, despite a net overall utility gain. Conversely, a policy based on increased levels of predicted performance, but incurring increased recruiting costs with a net overall utility loss, also may not be acceptable.

Examining the components of utility gains as well as overall net gain resulting from a policy change always provides a more comprehensive, understandable and credible basis of evaluating tradeoffs in a specific decision context. This is especially true when uncertainty increases over time, such as the impact on recruiting costs of changes in the attractiveness of military service in a contracting youth population. Examining individual outcomes also is important when the tradeoff rates among the components may not be consistent over the entire range of values. For instance, if overall utility is a function of both attrition rates and predicted performance gains, it is not possible to assign a single set of weights for both attrition and performance such that the weighted sum of the two will always yield the true value of the combination.

Comparisons of SD_y estimates are usually accomplished for jobs that have a sufficient number of supervisors and employees that permit consistency in estimates through averaging and for jobs in which expected variability of employees in dollar-valued performance is considered significant. Authors of comparative studies call for additional studies aimed at improving SD_y estimations by eliminating the source of inter-rater variability and rater confusion in making estimates. They also call for studies that may shed light on the accuracy of behaviorally based estimates by comparing them to objective, external measures of SD_y.

While the call for such comparative studies may make important methodological contributions, there are fewer calls for studies to report an overall productivity value of a selection policy in a true decision context. The next chapter examines a number of notable empirical utility analyses that provide such overall utility values within specific organizational settings.

CHAPTER 4. WHEN TESTING PAYS OFF: DOLLAR-VALUED EMPIRICAL RESULTS

A. THE UTILITY OF SELECTION PROGRAMS

Testing can save money because employees selected by valid tests are more productive than those selected by other methods. Use of Brogden's (1949) historic equation for estimating costs and benefits of a selection program has been a means of demonstrating that testing can save money. How much is saved depends on the predictive efficiency of the selection device, the selection ratio and two important recently applied situational variables--the variance of dollar-valued performance to the organization and costs associated with testing.

In earlier chapters, we reviewed and evaluated various interpretations of the validity coefficient, major utility analysis models, and alternative methods of estimating SD_y, the standard deviation of dollar-valued performance. The purpose of the present chapter is to describe productivity gains when all of the critical parameters in Brogden's equation are combined.

Despite the availability of Brogden's equation since 1949, utility analysis had not received widespread attention until recent years when practical means of estimating SD_y were developed. Using rational methods of estimating SD_y, Schmidt et al. (1979) applied utility analysis in the workplace to study computer programmers in the federal sector and Cascio and Silbey (1979) studied assessment centers in a sales organization. Both studies showed very large utilities: gains in the millions of dollars that can accrue to organizations annually from valid selection programs. The economic importance of scientific selection on work force productivity had not been fully recognized before the results of these early utility analyses became known.

By the 1980s, the number of utility analyses undertaken had increased. Results of these analyses also continued to show very large dollar gains that could be expected through selection, even when conservative values were used for all estimates. Today there appears to be a heightened awareness that utility analysis is not only a means of

possible to estimate PAT utilities in both the federal government and the economy as a whole, given an assumed testing cost for PAT of $10 per examinee as used in this study.

The study focused on the selection of computer programmers at the GS-5 through GS-9 grade levels, GS-5 being the lowest grade level in this occupational series. Beyond GS-9, it is unlikely that an aptitude test would be useful in selection. Applicants for higher level programmer positions are required to have considerable expertise in programming and are selected on the basis of experience and achievement. Most GS-9 programmers are promoted to GS-11 after one year. Similarly, most GS-5 grade level advance to GS-7 in one year and then to GS-9 the following year. Therefore, SD_y estimates were obtained for GS-9 through GS-11 grade levels.

Data from the Office of Personnel Management indicated that the total number of federal government computer programmers at all grade levels was 18,498 and the average yearly selection rate of GS-5 through GS-9 programmers was 618, with an average tenure of 9.69 years. Estimates based on U.S. census data of 1970 showed 166,556 programmer jobs for that year and that 10,210 new programmers could be hired each year in the U.S. economy. The 10,210 figure used in this study presumably is a substantial underestimate in view of the rapid expansion of this occupation in the 1980s.

Since it was not possible to determine prevailing selection ratios (SR) for computer programmers in the federal government or in the general economy, utilities were computed for SRs of 0.05 and intervals of 0.10 from 0.10 to 0.80.

The gains in utility or productivity as computed from Equation (2.6) are those that result when a valid procedure is introduced where previously no procedure or a totally invalid procedure has been used. The assumption that the true validity of the previous procedure is essentially zero may be true in some cases, but in other situations the PAT would, if introduced, replace a procedure with lower but non-zero true validity. The utilities were calculated assuming that previous procedure true validities ranged from 0.00 to 0.50.

Estimates of SD_y were provided by 105 experienced supervisors of computer programmers in 10 federal agencies using the questionnaire exhibited in the previous chapter describing the newly developed global estimation procedure.

Building all factors into Equation (2.6), the modified equation actually used in computing utilities:

$$\Delta U = tN_s(r_1 - r_2)SD_y\phi/p - N_s(C_1 - C_2)/p \quad (4.1)$$

where ΔU is the gain in productivity in dollars from using the new selection procedure for one year; t is the tenure in years of the average selectee (here 9.69); N_s is the number selected in a given year (this figure was 618 for federal government hires and 10,210 for hires in the U.S. economy); r_1 is the validity of the new procedure, here the PAT ($r_1 = 0.76$); r_2 is the validity of the previous procedure (r_2 ranges from 0 to 0.50); C_1 is the cost per applicant of the new procedure, here \$10; and C_2 is the cost per applicant of the previous procedure, here zero or \$10. The terms SD_y, ϕ and p are as defined previously in Equation (2.6), i.e., the standard deviation of performance in dollars, ordinate of the normal distribution at the cutoff score and the selection ratio, respectively. The figure for SD_y was the average of the two estimates obtained by using the global estimation procedure. Note that although this equation gives the productivity gain that results from substituting for one year the new (more valid) selection procedure for the previous procedure, not all of these gains are realized in the first year. They are spread out over the tenure of the new employees.

As mentioned in the previous chapter, the two estimates of SD_y were similar and thus the hypothesis that computer programmer productivity in dollars is normally distributed cannot be rejected. The average of the two estimates, \$10,413, was the SD_y value used in the utility calculations.

Table 4.1 shows the gain in productivity in millions of dollars that would result from one year's use of the PAT to select computer programmers in the federal government for different combinations of SR and previous procedure validity. At one extreme, if SR is 0.80 and the procedure PAT replaces has a validity of 0.50, the productivity gain is \$5.6 million for one year's use of the test. At the other extreme, when SR is 0.05 and the previous procedure has zero validity, the one year's test use productivity gain is \$97.2 million. (Again, the gains are not realized in the first year, but spread over the tenure of new employees.) The authors note that the figures in all cells of Table 4.1 are large--larger than most industrial and organizational psychologists would have expected.

Table 4.1. Estimated Productivity Increase from One Year's Use of the Programmer Aptitude Test to Select Computer Programmers in the Federal Government (In Millions of Dollars)

Selection ratio	True validity of previous procedures				
	0.00	0.20	0.30	0.40	0.50
0.05	97.2	71.7	58.9	46.1	33.3
0.10	82.8	60.1	50.1	39.2	28.3
0.20	66.0	48.6	40.0	31.3	22.6
0.30	54.7	40.3	33.1	25.9	18.7
0.40	45.6	34.6	27.6	21.6	15.6
0.50	37.6	27.7	22.8	17.8	12.9
0.60	30.4	22.4	18.4	14.4	10.4
0.70	23.4	17.2	14.1	11.1	8.0
0.80	16.5	12.2	10.0	7.8	5.6

Source: Schmidt et al. (1979), p. 622

Table 4.2 shows productivity gains for the economy as a whole resulting from use of the PAT or substitution of the PAT for less valid procedures. The figures in the table are based on the assumed yearly selection of 10,210 computer programmers nationwide and are for the total productivity gain. Gains per selectee in any cell of this table (and of the previous table) can be computed by dividing the cell entry by the number of selectees. As expected, the nationwide figures are considerably larger than the federal sector figures, exceeding $1 billion in several cells.

Schmidt et al. (1979) note that in addition to the assumption of linearity and normality discussed in Chapter 2, the productivity gains computed in this study are based on two additional assumptions. The first is the assumption that selection proceeds from the top-scoring applicant downward until the SR has been reached, the analysis assuming optimal selection procedures.

The second additional assumption is that all applicants offered the jobs accept them. The effect of rejecting hiring offers by applicants would be to increase the SR. For example, if a SR of 0.10 would yield the needed number of applicants, then if half of all job offers are rejected, the SR must be increased to 0.20, significantly reducing productivity gains.

Table 4.2. Estimated Productivity Increase from One Year's Use of the Programmer Aptitude Test to Select Computer Programmers in the U.S. Economy (In Millions of Dollars)

Selection ratio	True validity of previous procedures				
	0.00	0.20	0.30	0.40	0.50
0.05	1,605	1,184	973	761	550
0.10	1,367	1,008	828	648	468
0.20	1,091	804	661	517	373
0.30	903	666	547	428	309
0.40	753	555	455	356	257
0.50	622	459	376	295	213
0.60	501	370	304	238	172
0.70	387	285	234	183	132
0.80	273	201	165	129	93

Source: Schmidt et al. (1979), p. 623

An implicit assumption is also made that the organization's applicant pool is a representative sample of the potential applicant pool. An organization must be in a position to recruit and hire the most qualified applicants to obtain the full economic benefits of a valid selection procedure.

If the entire incumbent population of 18,498 programmers in the federal government had been selected by PAT with a validity of .76 in place of a procedure with a true validity of 0.30 and a SR of 0.20, then the productivity gain would have been about $1.2 billion. Expanding this example to the economy as a whole, the productivity gain that would have resulted is $10.78 billion. However, the productivity gains for an organization cannot be extrapolated in a simple way to all programmer jobs making up the national economy. If, nationwide, the number of people seeking programming jobs were to equal the number of jobs available, a zero-sum situation would arise. A fixed number of people would be allocated among various employers; for each employer hiring a superior applicant, another would hire an inferior applicant. Fortunately, the number of job seekers typically exceeds the number of jobs.

The authors further caution that productivity gains in individual jobs from improved selection cannot be extrapolated directly to productivity gains in the composite of all jobs making up the national economy. For example, if the potential gain economywide in the computer programmer occupation is $10.78 billion and if there is a total of *N* jobs in the economy (for programmers and all other jobs), the gain to be expected from use of improved selection procedures in all *N* jobs will not be as great as *N* times $10.78 billion. Since the total talent pool is not unlimited, gains due to selection in one job are partially offset by losses in other jobs. The size of the net gain for the economy depends on such factors as the number of jobs, the correlation among predicted job performance composites, and differences between jobs in SD_y.

Additionally, organizations must take into account such labor market characteristics as applicant supply and recruiting costs, making it very difficult to make projections of nationwide productivity gains from selection. Despite the difficulties in making estimates, net gains for the economy as a whole are clearly very large.

2. The Economic Impact of Predicting Job Performance of the Federal Work Force

In this study, Schmidt, Hunter, Outerbridge and Trattner (1986) evaluate the yearly productivity increases resulting from improved selection for white-collar jobs in the federal government in general rather than for a single white-collar job or occupation. In addition to being much broader in scope than earlier studies of selection utility, this study differs in several other advantageous respects.

a. Rather than estimating increases in predicted performance by use of linear regression-based decision theoretic equations, job performance differences between test-selected and non-test-selected employees could be determined empirically on the basis of direct job performance measurement.

b. Gains from improved selection could be expressed in a variety of ways: dollar-valued productivity increase; percentage increase in output; reduction in new hires and in the total work force when output is to remain constant; reduction in payroll costs produced by personnel costs; and selection gains expressed in terms of reduction in the proportion of "poor performers" among new hires and in the total work force.

c. The standard deviation of employee output could be estimated safely, if conservatively, as 20% of mean output, or alternatively, 40% of mean salary based on empirical findings described in the previous chapter.

Schmidt et al. (1986) were able to determine the means by which study participants had been selected for federal employment: by the Federal Service Entrance Examination (FSEE), a test of general mental ability including quantitative and verbal abilities; by the Professional and Administrative Career Examination (PACE) which superseded the FSEE; or by other means that usually did not include cognitive testing, such as internal promotion or upward mobility programs.

Job performance measures had been developed for use in three concurrent validation studies conducted earlier on Test 500, the written portion of the PACE described by McKillip, Trattner, Corts, and Wing (1977). The studies were conducted on Internal Revenue Service revenue officers (O'Leary & Trattner, 1977), customs inspectors (Corts, Muldrow, & Outerbridge, 1977), and social insurance claims examiners (Trattner, Corts, van Rijn, & Outerbridge, 1977).

In general, job performance was determined by a multimethod measurement approach that included work sample evaluations, job knowledge tests, and behaviorally anchored major duty rating scales completed by first-level supervisors. Employees' performance scores were the sum of the three standardized scores. The obtained difference in job performance between test- and non-test-selected employees, expressed in standard deviation units, was corrected for unreliability attributable to measurement error.

Accepting Hunter's (1983) finding that reliable cognitive tests such as the FSEE are comparably valid for all jobs of the required level of complexity of this study, the authors were able to use the average of the three estimated job performance differences to evaluate the effects of cognitive test selection vs. selection-by-other-means on the output of the federal white-collar work force as a whole.

Figures on federal white-collar civilian hiring were obtained from U.S. Office of Personnel Management (OPM) records for the five fiscal years 1977-1981 for each of the 18 general service (GS) levels. All categories of hires were converted to full-time equivalents (FTEs). Information was also available from which to compute average time from hiring to turnover, turnover rates by age intervals, and years of government service.

For each GS level, SD_y was estimated as 40% of the lowest salary for that grade, i.e., step 1 of 10 steps; thus SD_y estimates were conservative. Average gain in productivity in dollars per year per person hired resulting from the use of cognitive ability tests in selection was computed within each GS level as $d_t SD_y$ where d_t = mean difference in job performance in standard deviation units on a unit normal curve between the test-selected

and non-test-selected employee groups. The resulting figures were then adjusted for number of new hires and mean tenure in years of new hires. Cost of testing was not considered because the authors noted that costs were as high or higher for non-test-selection methods.

Percentage increase in output per year of new hires was computed as $20d_t$, where d_t is as defined above and 20 indicates 20% of mean output. The resulting percentage figure was used to compute the reduction in yearly hiring permitted by improved selection if output were to be held constant. "Poor performers" were defined as employees in the bottom 10% of performance among present employees. Reductions in poor performers were computed directly, using the value determined for d_t and the properties of the normal curve.

Table 4.3 shows the job performance differences between test-selected and non-test-selected employees. The observed job performance differences are quite similar across the three studies, ranging from 0.43 to 0.47 standard deviation units. After being corrected for unreliability in the measure of job performance, true performance difference ranged from 0.47 to 0.50, with a mean value of 0.487. This estimate, based on a total sample size of 673, should be reasonably stable. The standard error of the mean is only 0.007. The authors point out these findings cannot be explained on the basis of length of time on the job or differences between the two applicant pools.

Table 4.4 shows the number of new full-time equivalents by GS grade level for the five fiscal years studied, the lowest salary level (out of 10 levels for each GS level) and the estimate of SD_y obtained by the 40% of salary proportional rule.

Table 4.5 shows the dollar-valued productivity increases that result from selecting new hires through tests. Total expected gain from one year's use is over $7.8 billion; when the selection procedure is used for a decade, the projected gain is substantially larger--over $78 billion.

In practice, most people hired at the GS 13-18 levels are experienced managers or professionals, e.g., lawyers and scientists; ability testing is not used in the hiring process for these individuals. If cognitive testing were used only for hiring at the GS 1-4 and GS 5-8 ranges, the dollar value of productivity increases from one year's test use would be about $6.2 billion. This is 79% of the value obtained when tests are used for one year of hiring at all GS levels ($7.8 billion).

Table 4.3. Job Performance Difference Between Test-Selected and Non-Test-Selected Employee in Three Occupations

Occupation	*N*	Mean tenure in job (years)	Mean job perform-ance [a]	Within group *SD*	Total group *SD*	Difference in total group *SD* units[b]	Criterion reliability	True difference in total group *SD* units[c]
1. *Internal Revenue Officers*								
(O'Leary & Trattner, 1977)								
Test-selected	136	7.47	0.50	2.19	2.19	0.43	0.85	0.47
Selected without test	156	8.15	−0.43	2.11				
2. *Customs Inspectors*								
(Corts, Muldrow, & Outerbridge, 1977)								
Test-selected	99	5.15	0.55	2.63	2.57	0.47	0.90	0.50
Selected without test	79	5.00	−0.65	2.36				
3. *Social Insurance Claims Examiners*								
(Trattner, Corts, van Rijn, & Outerbridge, 1977)								
Test-selected	112	6.64	0.68	2.88	2.88	0.46	0.90	0.49
Selected without test	91	8.35	−0.63	2.82				

Source: Schmidt et al. (1986), p. 9.

[a] Based on a composite measure of work-sample scores, job-information scores, and supervisory ratings and/or rankings.

[b] All values significant beyond 0.001 level.

[c] The mean value across the three occupations is 0.478.

Table 4.4. Number of New White-Collar Hires in the Federal Government for Five Recent Years, Salaries, and Estimated Standard Deviations of Job Performance in Dollars

General schedule (GS) level	Full-time equivalents hired in fiscal year					Mean no. FTEs hired	Lowest 1984	Estimated
	1977	1978	1979	1980	1981	1977-81	salary [a]	SD_y [b]
1	6,561	7,085	6,204	8,460	11,843	8,031	$9,023	$3,609
2	39,520	42,021	36,227	40,102	33,744	38,335	10,146	4,058
3	69,940	69,238	54,604	68,675	60,512	64,594	11,070	4,428
4	44,975	48,996	38,493	55,069	45,924	46,691	12,427	4,971
5	28,358	31,337	22,315	30,472	22,598	27,016	13,903	5,561
6	3,352	3,715	3,361	4,100	3,381	3,672	15,497	6,199
7	13,812	15,040	14,130	16,354	12,921	14,451	17,221	6,888
8	612	567	385	477	609	530	19,073	7,629
9	7,856	9,511	8,213	10,138	8,838	8,911	21,066	8,426
10	293	282	281	312	413	316	23,199	9,280
11	5,206	5,653	5,279	6,456	5,190	5,557	25,489	10,196
12	3,257	3,591	3,412	4,495	3,692	3,689	30,549	12,220
13	1,711	1,681	1,796	2,484	1,859	1,906	36,327	14,531
14	880	879	980	1,280	1,167	1,037	42,928	17,171
15	778	726	658	1,077	955	839	50,495	20,198
16	113	97	98	67	48	85	59,223	23,689
17	84	50	50	22	25	46	69,376	27,750
18	52	38	23	2	11	25	81,311	32,524
	Total					225,731	—	—
	Means						13,573 [c]	5,429 [c]
	SDs						5,627 [c]	2,251 [c]

Source: Schmidt et al. (1986), p. 10.

a There are 10 salary steps within each GS level. New hires begin at the lowest step (Step 1), which is reported here. They advanced 1 step per year for the first 3 years. 1 step every two years for the next 3 steps or 6 years, and 1 step every 3 years for the next 9 years. Hence the salary figures reported here are very conservative.

b 40% of salary; see text for further explanation.

c Weighted by number hired at each GS level.

Table 4.5. Dollar Value of Productivity Increases in the Federal Government Resulting from the Use of Valid Cognitive Ability Selection Tests for 18 Job Levels

General schedule (GS) level	(1) ΔU/hire/year = 0.487 SDy [a]	(2) ΔU/year = N_S (0.487) SDy [b,c]	(3) Mean weighted tenure in years	Dollar value of productivity gains for selected durations of test use in years[c] 1 yr	5 yrs	10 yrs
1	$1,758	14.12	13.01	183.70	918.51	1,837.01
2	1,976	75.75	12.92	978.69	4,893.45	9,786.90
3	2,156	139.26	13.05	1,817.34	9,086.72	18,173.43
4	2,421	113.04	12.92	1,460.48	7,302.38	14,604.77
5	2,708	73.16	13.16	962.79	4,813.93	9,627.86
6	3,019	11.08	13.24	146.70	733.50	1,466.99
7	3,354	48.47	13.42	650.47	3,252.34	6,504.67
8	3,715	1.97	13.17	25.95	129.72	259.45
9	4,103	36.56	13.47	492.46	2,462.32	4,924.63
10	4,519	1.43	13.26	18.96	94.81	189.62
11	4,965	27.59	13.56	374.12	1,870.60	3,741.20
12	5,951	21.95	13.58	298.08	1,490.41	2,980.81
13	7,076	13.49	13.52	182.38	911.92	1,823.85
14	8,362	8.67	13.25	114.88	574.39	1,148.77
15	9,836	8.25	12.92	106.59	532.95	1,065.90
16	11,536	.98	12.97	12.71	63.55	127.10
17	13,514	.62	13.01	8.07	40.33	80.66
18	15,839	.40	12.84	5.14	25.68	51.36
Totals		596.79		7,839.51	39,197.51	78,394.98

Source: Schmidt et al. (1986), p. 12.

[a] In 1984 dollars.

[b] Corresponds to mean job tenure of 1 year for new hires.

[c] In millions of dollars.

Authors of this study note that in interpreting the dollar utility figures in Table 4.3, the tenure figures used in the utility calculations ignore promotions and career advancement. If typical advancements were considered, appropriate SD_y values would increase, increasing actual productivity gains from selection. As a consequence of assuming employees remain in their GS entry level for the duration of their employment, utility figures were again underestimated.

Schmidt et al. state that these utility values will appear to some readers to be overly large and even implausible. One approach to demonstrating the realistic nature of these values is to express them as percentage increase in output:

$$\text{percent increase/selectee/year} = 0.487(20) = 9.74\% \quad .$$

Thus, the output of test-selected employees averages 9.74% higher than that of non-test-selected employees; alternatively, the output of test-selected employees averages 109.74% of that of non-test-selected employees. It is this difference in output that produces the dollar gains shown in Table 4.3.

Some organizations, the authors note, may wish to maintain output at some constant, fixed level while reducing the costs of producing that level of output. Improved selection allows such organizations to reduce the size of the work force gradually to produce the desired fixed level of output. In this study, since test-selected employees have been shown to have 9.74% greater output, only 91.12% (100/1.0974) as many need to be hired as would be needed if new hires were non-test selected.

Table 4.6 shows the expected reduction in the number of new hires necessary to maintain constant output by a transition from non-test selection methods to cognitive tests. Column 5 of the table gives the yearly payroll savings produced by the reduction or hiring. Test selection for the GS 1-4 and GS 5-8 levels alone would result in a yearly hiring reduction of 18,056, for a per-year savings payroll cost of $217.39 million. Taking expected tenure of these employees into account, the figure rises to $2.8 billion. This analysis can be applied not only to reduction in new hires but also to reduction in the size of the total workforce.

If the percentage of "poor performers" is arbitrarily set at the 10th percentile in the performance distribution of non-test-selected employees, and the effect of using cognitive selection tests is to raise the mean level of job performance by 0.487 standard deviation units, then the expected percentage in the poor performer range among test-selected

Table 4.6. Reduction in the Number of Yearly New FTE Hires Necessary to Maintain Constant Output and Resultant Yearly Reductions in Payroll Costs (One Year's Test Use)

	General schedule (GS) level	(1) No of new FTE hires required without use of valid tests	(2) No. of new FTE hires required with use of valid tests	(3) Reduction in the yearly number hired	(4) Mean yearly salary [a]	(5) Yearly payroll savings[b]
1	GS 1-4	$157,651	143,651	14,000	$11,143	156.00
2	GS 5-8	45,669	41,613	4,056	15,141	61.41
3	GS 9-12	18,473	16,832	1,641	24,327	39.92
4	GS 13-18	3,938	3,588	350	42,250	14.79
5	Sum of (1) through (4) above	225,731	205,686	20,045	13,573	272.07
6	Sum of (1) and (2) above	203,320	185,264	18,056	12,040	217.39
7	Sum of (1) and (2) above plus 25% of (3) and (4)	208,922	190,370	18,552	12,454	231.05

Source: Schmidt et al. (1986), p. 17.

a Frequency weighted mean; in dollars.

b In millions of (1984) dollars; frequency weighted totals based on lowest 1984 salaries for each GS level; employee benefits (which amount to about 14% of salary) are not included.

employees is 3.86%, a reduction of 61.4%. Across all GS levels, use of cognitive selection tests would be expected to lead to 13,861 fewer poor performers coming onto the federal payroll per year.

Schmidt et al. (1986) point out that the results of the study "depend on the difference of 0.487 standard deviation units in job performance between test-selected and non-test-selected employees" (p. 19). They give three reasons why this difference may be an underestimate: performance-related information on current employees not available for use with outside applicants was included; many current employees were initially selected into lower level jobs from the outside using cognitive tests; and procedures used to select specifically for the jobs in the present study sometimes included as one component scores on other, less difficult, cognitive tests. The authors computed a theoretical estimate of the job performance difference and found that the observed estimate is about 19% smaller than the theoretical estimate, confirming their belief that the empirical data probably somewhat underestimates the difference in job performance.

The authors caution that the productivity gains presented in this study represent the differences to be expected between a federal system that uses ability tests to select for all jobs and a federal system that uses only traditional evaluations of education and experience to select for all jobs. Neither of these hypothetical systems describes the current federal selection systems. An unknown percentage of new federal hires, especially at the GS 2-5 levels, are currently selected based on ability tests of one kind or another. As a purely hypothetical example, if 30% of new hires at all GS levels were currently selected using cognitive tests, the potential productivity gains available to the government would be 70% of those reported in this study. The authors conclude that although the exact potential gains are unknown, it is virtually certain that they are very large.

Although the precise percentage of new federal hires currently selected based on ability tests of one kind or another is unknown, reports on test usage by the Civil Service Commission (Wing, 1977; Campbell, 1979) provide data relevant to the hiring period reported in the study. In 1978, 63 different standardized tests were being used singly or in combination to fill entry-level positions in about 300 occupations in which extensive experience is neither expected nor required. Data indicate that six major written tests were given to 700,000 applicants yearly for clerical and lower-level jobs and for entry-level professional and administrative occupations. About half of this total number of applicants for federal positions seek the jobs covered by these examinations.

Further, on the basis of a suit brought against the Office of Personnel Management on the grounds that PACE had an adverse impact on minorities, the government entered into a consent decree in 1981 agreeing to eliminate the PACE which had been administered to nearly 200,000 applicants yearly.

Considering the figures just mentioned and the fact that most individuals hired at the GS-11 level or higher are not tested, the "purely" hypothetical" example given by the authors of 30% new hires currently being tested might be a reasonable ballpark estimate. If this is the case, then only 30% of the productivity gains reported in this study may actually be realized in practice.

Further, Hunter (1983) shows that the level of validity for general mental ability tests depends on the complexity of information processing demands imposed by the job. He found validities ranging from 0.56 for the highest level of complexity jobs to 0.23 for the lowest level of complexity jobs. Using the average difference in test-selected and non-test-selected employees for three jobs relatively close in complexity does not appear to represent adequately the complete range of complexity of jobs in the federal white-collar work force as a whole. Since the largest number of employees serve in jobs of lower levels of complexity at lower grade levels, differential estimates of job performance differences, weighted across all grade levels, may be significantly smaller than the across-the-board average difference value used. Consequently, using more realistic, smaller estimates of job performance differences may result in considerably smaller figures of productivity gains.

3. Productivity Gains by a Hierarchical Model of Talent Allocation

Hunter and Schmidt (1982) were interested in evaluating the impact of personnel classification to jobs on national productivity. The relevant focus of classification of personnel to multiple jobs is no longer the selection model for a single job, but the personnel classification model that assumes each applicant will be assigned to one of several possible jobs. The classification model assigns individuals to jobs in such a way as to maximize overall productivity, while ensuring that each job receives the required number of employees. Classification may use a separate equation for predicting performance for each job (Brogden, 1955, 1964).

In Brogden's (1959) classical study, making a number of simplifying assumptions, he showed that the general solution of average productivity is:

$$U = v\sqrt{1 - r}\ U_o \tag{4.2}$$

where v is the actual validity, r is the correlation among predicted performance scores, and U_o is expected value (in standard units) of predicted performance for relevant selection ratios, assuming 9 jobs, zero correlations among predictors of performance and the validity is 1.00.

As an example of the productivity implications of Brogden's (1959) findings, Hunter and Schmidt (1982) consider an economy in which:

> (1) there are only 10 different jobs (i.e., 10 unique regression equations for predicting job performance); (2) yearly SD_y is \$7,000 for all jobs; (3) validity is .45 for all jobs, (4) the average correlation among prediction composites (mean $r_{y_i\ y_j}$) is .85; and (5) the labor force is 90 million strong. If we further assume that every member of the labor force will be assigned a job (i.e., there is no reject category), . . . that the mean standard job performance score when validity = 1.00 and $r_{y_i\ y_j} = 0$ is 1.54. The difference in yearly productivity between random assignment of the 90 million workers to jobs and assignment based on the classification model is then:
>
> $$\Delta U = .45\sqrt{1 - .85}\ (1.54)(\$7{,}000)(90{,}000{,}000)$$
>
> $$\Delta U = 169 \text{ billion dollars.}$$
>
> Obviously, the productivity implications of appropriate ability-job requirement matching can be substantial. This figure is, of course, constrained by Brogden's somewhat unrealistic simplifying assumptions (e.g., SD_y, mean productivity, and number of incumbents are assumed the same for all jobs). (p. 261)

Hunter and Schmidt developed an hierarchical model of talent allocation that contrasts a society in which all jobs are allocated on the basis of a few cognitive abilities to a society in which all jobs are assigned randomly to members of the labor force. A four-class categorization scheme was chosen: management-professional; skilled trades (including crafts, such as bricklaying, as well as industrially defined trades, such as tool and die making); clerical (here the authors actually mean all white-collar work at a non-managerial level); and semiskilled and unskilled labor (the residual blue-collar and farm-labor workers). In terms of ability correlates, four distinct "kinds" of jobs are assumed to exist in the economy.

The proportion of the labor force in each of these categories based on U.S. Bureau of the Census (1977) figures is 24, 12, 24, and 40%, respectively. For the authors' preliminary calculations, they have taken mean output to be equal to median income in these groups. The Census Bureau places median 1976 incomes at $12,818, $11,476, $6,668, and $4,883 for these four groups, respectively.

For any given job the authors write:

$$y = \mu + r_{xy} SD_y \bar{Z}_x + e$$

where

$\bar{Z}_x$ is ability expressed in standard score units (mean O, standard deviation = 1),

y is individual performance on the job expressed in dollars,

μ is the mean performance in dollars of individual selected to the job *without* use of the test,

SD_y is the performance standard deviation in dollars of persons selected to the job *without* use of the test,

r_{xy} is the population correlation between ability and performance (for the entire working population), and

e is the residual error of prediction. (p. 262)

If a group of persons is selected on the basis of ability, and if the mean ability of that group is given by $\bar{Z}_x$, then the mean performance, y, is given by $y = \mu + r_{xy} SD_y \bar{Z}_x$. This equation differs from earlier equations for ΔU/selectee in that it includes the term μ. It gives the mean absolute level of productivity rather than the increment in productivity (marginal utility). It omits the term for testing costs that are considered by the authors as negligible compared to utility gains.

For random selection, mean ability of those selected is the same as the mean for the population as a whole, which is zero if ability is expressed in standard scores. Thus, for random selection, mean productivity for a given group is simply given by the constant μ for that group, which is the mean output assumed earlier (i.e., $12,818 for the managerial-professional group, $11,476 for the skilled-trades group, etc.). The mean output for the country as a whole is the weighted average of these means, where each group is weighted by the number of persons in that group--$8,007 per year.

The model the authors employed to determine the impact of job assignment strategies on productivity was a multi-ordered selection process; organizations hiring people for managerial-professional positions have first choice of workers because workers prefer these highest paid jobs. There is then a similar selection for the skilled trades for those who do not land a job in the top paying category and the process is the same for the clerical and unskilled labor categories. It was assumed by the authors that each successive job category selects its workers from those remaining after the previous category has attracted those it requires. No reject category is used, i.e., everyone must be assigned to a job. Within these constraints, the job allocation model can be either univariate or multivariate.

In univariate selection, job assignments are all made on the same ability, with gains attributed to selection for one job being partially offset by losses attributed to the same selection process to other jobs. However, the authors state this cancellation effect will not be complete unless $r_{xy_i} SD_{y_i}$ is equal for all jobs. The univariate model requires that the labor force be broken into separate categories on that test score; the top 24% who go to managerial and professional jobs, the next 12% go to skilled trades, etc. For purposes of comparison, this model is called the "univariate selection" model. For this model, there is a maximum of counterbalancing between the gains produced by selecting the brightest for managerial-professional jobs, and the losses produced by selecting the dullest for unskilled labor. However, because the authors' review of utility studies suggests that individual differences in output in dollars in high-paying jobs (i.e., absolute values of SD_y) are greater than such differences in lower-paying jobs, the gains at the top will be larger than the losses at the bottom. Thus, because the model assumes that the standard deviation in dollar output is proportional to mean dollar output, their model predicts that univariate selection will yield higher utility than will random selection.

The optimal prediction of job performance requires different ability combinations for different jobs. The authors note that their multivariate models crudely but faithfully preserve these distinctions.

The assumptions of the abilities required for each job category and their correlational relationships are shown in path analytic form in Figure 4.1.

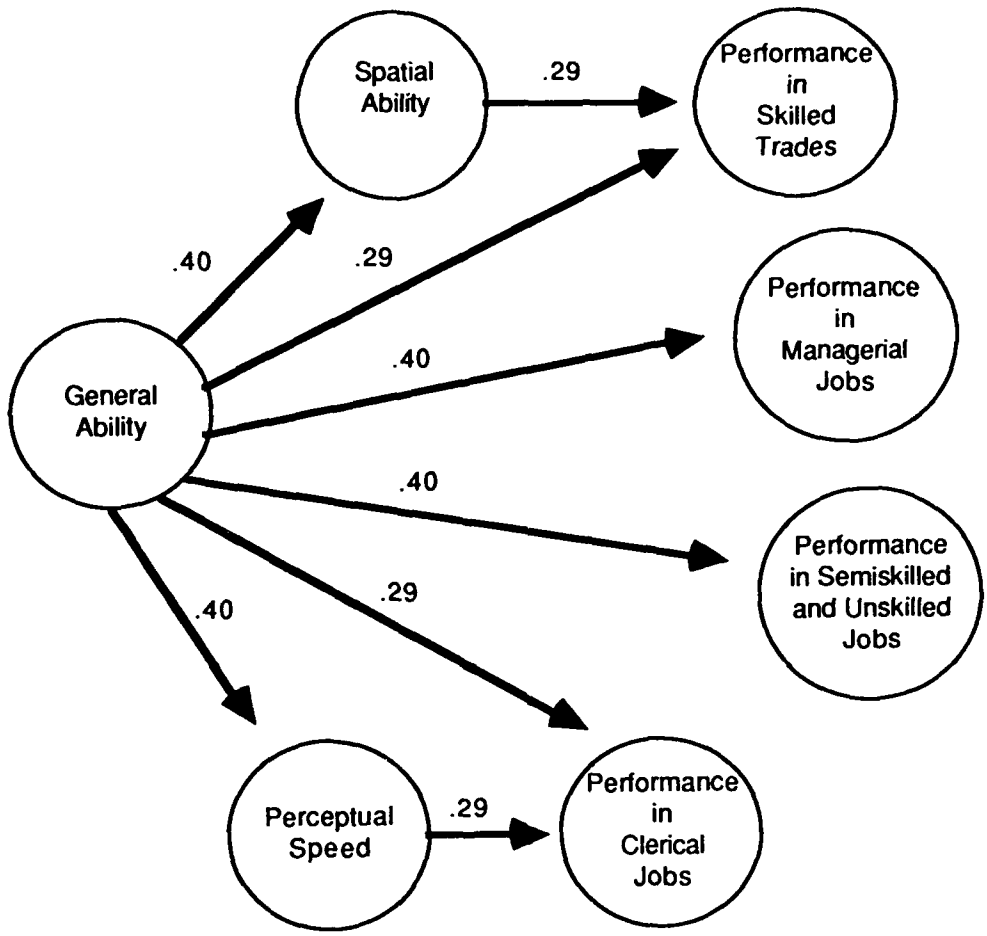

Figure 4.1. The Correlational Assumptions in the Multivariate Selection Model in Path Analytic Form

Source: Hunter and Schmidt (1982), p. 265.

The authors note that if different abilities are required in different jobs, then to the extent that those abilities are less than perfectly correlated, multivariate selection (i.e., selection on combined ability test scores) will be less prone to losses due to selection; that is, gains from selecting high-ability people for one job will be offset less by selection of low-ability people to other jobs than in the case of univariate selection (Brogden, 1959). Thus the model predicts greater gains in overall utility for multivariate selection than for univariate selection.

Table 4.7 shows the mean output for employees assigned to jobs by means of random, univariate and multivariate selection. The data were computed under the most conservative assumption about dollar-valued job performance: that SD_y is only 16% of the mean output. Schmidt and Hunter (1982) reported this value to be greatly underestimated because only partial performance dimensions, e.g., tenure or training costs, were used to compute the mean output in the studies surveyed.

Table 4.7. Mean Annual Output of Workers in 1976 Dollars in Four Occupational Categories with Three Different Personnel Assignment Strategies and SD_y = 0.16u

Occupation Class	No. of Persons[a]	Percent	Random Selection	Univariate Selection	Multivariate Selection
Professional/ managerial	25,085	24	12,818	13,815	13,815
Skilled	12,593	12	11,476	11,766	12,137
Clerical	25,040	24	6,668	6,627	6,847
Semiskilled and Unskilled	41,605	40	4,883	4,549	4,625
Total	104,323	100	8,007	8,137	8,265

Source: Hunter and Schmidt (1982), p. 266

[a] In thousands.

As noted by the authors, the mean output in goods and services, if people were randomly assigned to jobs, would be $8,007 per person per year. Apparently, the least able individual is more productive as a professional/manager than he is as a semiskilled/unskilled worker, i.e., 62% more productive. However, if persons were assigned on the basis of general ability, then average production would be $8,137 per year per person. This figure is a compromise between the increase in production for those in high-output jobs and the decrease in average production for those in low-output jobs (i.e., because better than average workers are assigned to the high-output jobs, the resulting average output improves). By the same token, lower than average workers are left to do the lower-output jobs and, hence, average output on those jobs decreases. Assignment on general ability provides an increase of $130 per person per year, aggregated across 104 million workers, means a difference of $13.5 billion in the productivity of the nation as a whole.

If multivariate selection is used, then the mean output rises to $8,265 per person per year, or an increase of $26.8 billion per year, even under the most conservative assumptions. The authors give a number of reasons why these figures are extremely conservative: use of an extremely low SD_y value (16%); inclusion of too few abilities and job categories; use of lower than possible validities, and use of census mean income figures universally recognized as underestimates.

To provide a more accurate, though still conservative estimate of the impact of selection, the authors set SD_y at 40% of mean salary and adjusted all incomes upward by 10% to allow for underreporting of income and by 20% to allow for inflation since 1976. The figures resulting from these adjustments are shown in Table 4.8. Under these assumptions, the authors point out, the average productivity difference between random selection and univariate selection is approximately $423 per worker per year ($10,832-10,409), or 44.1 billion dollars per year for the labor force as a whole. Similarly, the difference between random and multivariate selection is $839 per worker per year ($11,248-10,409), or $87.5 billion per year economywide. If SD_y is taken as 70% of mean salary, these figures are $76.9 billion and $152.6 billion per year, respectively. These results are summarized in Table 4.9.

Hunter and Schmidt's model assumes that mean output is equal to mean income, but in 1970, wages were only 57% of output. Thus, for 1980, all output figures reported in Tables 4.8 and 4.9 may be multiplied by 1.75 (1/0.57) to obtain more realistic productivity gain estimates.

Table 4.8. Mean Annual Output of Workers in 1980 Dollars in Four Occupational Categories with Three Different Personnel Assignment Strategies and SD_y = 0.40u

Occupation Class	No. of Persons[a]	Percent	Random Selection	Univariate Selection	Multivariate Selection
Professional/ managerial	25,085	24	16,663	19,902	19,902
Skilled	12,593	12	14,919	15,859	17,065
Clerical	25,040	24	8,668	8,534	9,251
Semiskilled and Unskilled	41,605	40	6,348	5,262	5,508
Total	104,323	100	10,409	10,832	11,248

Source: Hunter and Schmidt (1982), p. 268

[a] In thousands.

Table 4.9. Estima' Productivity Differences Between Selection Strategies (In Billions of Dollars)

	SD_y as Percent of Salary		
	16%	40%[a]	70%[a]
Univariate vs. Random	13.5	44.1	76.9
Multivariate vs. Random	26.8	87.5	152.6
Multivariate vs. Univariate	13.3	43.3	54.2

Source: Hunter and Schmidt (1987), p. 269

[a] 1976 salaries adjusted upward by 10% to allow for underreporting and by 20% to allow for inflation between 1976 and 1980.

The authors conclude that the real-world implications for productivity improvements using a multivariate selection model is a reasonable approximation to an optimal procedure for job assignment in the U.S. economy. They note that their analysis of job allocation procedures on national productivity may be the first of its kind but is neither definitive nor complete. As the authors also note, an interdisciplinary team including economists would be required to incorporate proper economic cost considerations and labor force dynamics needed to refine the model and enhance its credibility.

The authors state that despite the limitations in the analysis, the way in which talent is allocated to jobs in the economy does have a significant impact on national productivity. By moving from current employment decisions which are certainly better than those that would follow random selection from applicant pools, productivity gains can be conservatively estimated between $43 and $54 billion per year. Chapter 9 provides additional discussion of this study.

4. Utility of an Assessment Center as a Selection Procedure

Burke and Frederick (1986), using an interdisciplinary approach, compared utility estimates for an assessment center that was used to select district managers at a large national manufacturing organization. As described in the previous chapter, two consensus-seeking procedures, the global estimation procedure, and 40% and 70% of mean salary were used for estimating SD_y.

The authors incorporated Boudreau's (1983b) economic concepts (variable costs, taxes, and discounting) into the general utility equation. Additionally, the 1986 study used Boudreau's (1983a) definition of the payoff function as net benefits, the difference between sales value (e.g., sales revenue) and service costs (e.g., salary and benefits). Burke and Frederick note that while others define the payoff function as the "value of output as sold," neither definition represents the "correct" one, but for many purposes Boudreau's payoff function offers a more complete expression.

The present utility analysis evaluated an assessment center that had been in operation for seven years and had assessed 132 area sales representatives, 29 of whom were selected at the time of the study to fill district sales managerial position. Thus, the selection ratio was 0.22.

Boudreau's (1983a) utility formula was used in this analysis:

$$\Delta U = (N_s)\left[\sum_{t=1}^{T}\left(1/(1+i)^t\right)(SD_y)(1+V)\times(1-TAX)(\hat{p}_x r_y)(\bar{Z}_s)\right] - C(1-TAX) \quad ,$$

where

ΔU = total dollar value of personnel program after variable costs, taxes, and discounting;

N_s = the number of employees selected;

t = the time period in which a productivity increase occurs;

i = the discount rate;

SD_y = the standard deviation of job performance in dollars (this value is comparable to Boudreau's SD_{sv}, standard deviation of the sales value of productivity);

V = the proportion of SD_y represented by variable costs;

TAX = the organization's applicable tax rate;

$\hat{p}_x r_y$ = the estimated correlation between predictor x and true scores on the criterion y in the population, the true operational validity coefficient;

$\bar{Z}_s$ = the mean standard score on the predictor of those selected; and

C = the cost of the personnel program.

For purposes of this study, the company's discount (interest) rate of 25% adjusted for inflation was used in utility calculations. Using the average tenure of a district sales manager of 4.2 years, the average inflation rate for this time period was 7.2%. On the basis of the average tenure of a district sales manager, it was assumed that four years was the period of the assessment center's effects. Thus, the present value of the assessment center was calculated on the basis of a duration of four years and an average discount rate of 17.8%.

The various procedures for estimating SD_y in the present study were described in the previous chapter. The value for variable costs (V) considers the proportion of dollar sales value compared to the operating costs. An average of V across five zones of sales achievement was found to be 4.8%, and because there was a positive relationship between combined operating costs and sales volume, a value of −0.048 was used in subsequent

utility computations. The applicable corporate tax rate of 0.49 (49%) was used for *TAX*. Both *V* and *TAX* were assumed to be constant for each time period.

A multiple correlation of 0.61 was computed between a composite assessment center score and an overall performance rating. An estimated population cross-validated value of 0.41 was determined by using Cattin's (1980) formula. Correction for restriction in range raised this correlation to 0.46. Assuming the reliability of the criteria measure to be 0.6, based on validity generalization findings of Pearlman, Schmidt, and Hunter (1980), the resulting true operational validity coefficient, $\hat{\rho}_x r_y$, was found to be 0.59.

The mean $\overline{Z}$ score on the predictor of those selected was calculated to be 0.872. This value differs from the mean $\overline{Z}$ score of 1.3 obtained on the basis of the selection ratio of 0.22, assuming a strict top-down selection and using normal distribution tables.

The cost of the assessment center, including costs in establishing the center, and costs of consultants, assessors, and candidates, was computed to be $263,636; the cost of assessing one individual was $2,000, spreading costs over 132 assessed individuals. On the basis of experience in selecting 29 individuals, the cost of selecting one district sales manager was $9,091. The authors note that the value of $403.28 was reported by Cascio and Silbey (1979) for assessing comparable second-level managers. They attribute the higher cost in the present study to inclusion of consulting fees to maintain the center and other higher cost estimates provided by the cost accounting department.

The per-selectee utility of the assessment center was calculated for each of eight types of SD_y estimates. Since a prior selection procedure was in place that used interviews, eight sets of utility estimates were calculated that incorporated the validity and cost of the prior interview procedure. Because information on components of the interviewing program was unavailable, data on cost of interviewing (Cascio & Silbey, 1979) and on the true validity of interviews (Hunter & Hunter, 1984) were used to adjust estimated per-selectee gains. The total estimated cost to select 29 district sales managers by means of an interviewing program was computed to be $50,485; the true operational validity for the interview was estimated to be 0.16. These values were used in a modified form of Equation (4.2) to adjust the validity and cost of the assessment center over random selection, thereby assisting in computing refined estimates.

The unadjusted and adjusted utility estimates were compared to the estimated utility gains that would have resulted from top-down selection. Assuming that the selection ratio and all other factors were to remain constant, top-down selection would result in a mean $\overline{Z}$

score on the predictor of 1.29 from the present group of assessees. Using a mean $\bar{Z}$ score of 1.29, subsequent utility gains were then computed.

Table 4.10 shows the estimated per-selectee and total (present value) assessment payoff. The payoff varies somewhat depending on the type of SD_y estimate used and whether or not a previous selection procedure is incorporated into the estimates. As noted in the previous chapter, Burke and Frederick state that a tentative case can be made for using the SD_y estimate based on four percentile points individual feedback (Procedure B) and on global estimation (Schmidt et al., 1979), both of which produced virtually identical results.

The results clearly show that regardless of the SD_y estimating procedure used, the assessment center is a cost effective means of selecting sales managers. The estimated per-selectee gain for a four-year period over random selection for global estimation is $17,143, and the gain over selection based on interviewing is $12,124. The net productivity gain in assessment center selection of 29 district managers is $497,150 over random selection and $351,616 over selection by interviewing. Extending these findings, if 100 district managers were to be selected by assessment center means instead of interviews, the four-year gain would be over $1.2 million. The estimating procedure only partially reflects employee flows (Boudreau, 1983b), which, the authors note, are very likely to increase the net utility gains.

The potential value of assessment center selection becomes more clearly evident when estimated utilities are compared using the current hiring practice versus a top-down selection procedure. Table 4.10 shows that utility gains are substantially reduced by selecting managers with an average predictor score of 0.872 standard deviations above the mean (current practice) over the top-down selection procedure. For example, estimated adjusted total utility gains, using four percentile points global estimates, is $351,614 for the current strategy compared to $576,420 for the top-down strategy; only 62% of the potential gain is retained under the current practice.

In discussions with management personnel, the authors found that managers were unaware of the economic impact of the assessment center selection program. Even the highest estimates of utility gains were well accepted by corporate management. Credibility is attributed to such factors as the use of realistic and possibly conservative assumptions,

Table 4.10. Assessment Center Utility Analysis Results in Dollars

SD_y estimation procedure	Estimated per selectee gain (for a 4-year period)			Estimated total utility gain (for a 4-year period)			
	SD_y	Unadjusted	Adjusted[a]	Unadjusted	% top-down[b]	Adjusted	% top-down[b]
Based on 15th, 50th and 85th percentile points							
Procedure A	27,500	13,914	9,771	403,506	61	283,367	60
Procedure B	28,151	14,353	10,091	416,241	61	292,648	60
Schmidt, Hunter, McKenzie and Muldrow (1979)	35,192	19,103	13,552	553,979	63	393,033	62
Based on 15th, 85th and 97th percentile points							
Procedure A	38,333	21,222	15,097	615,423	63	437,815	63
Procedure B	32,323	17,167	12,142	497,855	62	352,128	61
Schmidt, Hunter, McKenzie and Muldrow (1979)	32,287	17,143	12,124	497,150	62	351,616	61
Percentage of mean salary							
40% of mean salary	12,789	3,991	2,538	115,727	49	73,629	46
70% of mean salary	22,381	10,461	7,254	303,367	59	210,384	60

Source: Burke and Frederick (1986), p. 337.

[a] Adjusted figures are assessment center utility results adjusted for the estimated validity and cost of the interviewing program.

[b] For top-down selection, the mean z score on the predictor was 1.29 for the top 29 assessment scores (the selection ratio was 0.22).

the employment of an interdisciplinary approach incorporating economic considerations, and the involvement of the accounting, sales, corporate strategic planning and tax departments.

Burke and Frederick's (1986) study is an outstanding utility analysis conducted in a realistic situational context. The major purpose of the study was to demonstrate the utility of an existing assessment center selection program in a particular organization, rather than the use of utility analysis as a tool to choose among selection alternatives or propose modifications. It fully achieves its purpose by clearly specifying assumptions, defining the components used in the analysis, and providing accurate or conservative estimates. The organization for which this analysis was done has a more comprehensive understanding of the economic implications of assessment center selection of its district sales managers.

5. Selection Utility for U.S. Park Rangers

Schmidt, Mack and Hunter's (1984) study evaluates the utility of a testing procedure versus an unstructured employment interview in the selection of park rangers in the U.S. Park Service. The study also evaluates the impact on employee output of three modes of selection test use: top-down selection; minimum required test scores equal to the mean; and minimum score at one standard deviation below the mean.

Estimates of SD_y were obtained from 114 first-line supervisors of park rangers sampled from various national parks within the National Park Service of the U.S. Department of the Interior.

Supervisors were asked to estimate the dollar value of the yearly services of superior (85th percentile), average (50th percentile), and below average (15th percentile) entry-level park rangers following the global estimation procedure developed by Schmidt et al. (1979).

Hunter's (1983) validity generalization results of the General Aptitude Test Battery's general mental ability composite were used to estimate the true validity of the park ranger test selection procedure. The job of entry-level park ranger was judged to fall within a middle level of complexity with a true validity of about 0.51. The battery actually used for park ranger selection was the Professional and Administrative Career Examination (PACE), a multifaceted measure of general mental ability. As previously noted, criterion-related validity studies for three jobs were conducted. Levels of validity for the PACE similar to that provided by Hunter's (1983) meta-analysis were found in the studies.

The selection ratio for the PACE averaged about .10 over the years of its use (1974-1982). The test was used to select new employees for a variety of entry-level administrative and professional occupations. The average number of entry-level park rangers hired per year during the 1978-1981 period using the PACE, according to figures supplied by the U.S. Park Service, was 83. The standard deviation across these years was 42.11. Because of this variability in number of subjects, utility was evaluated not only for the rounded mean figure of 80 but also for a range of values above and below that figure.

Since the time the PACE was discontinued in the early 1980s, for reasons noted earlier, entry-level park rangers were hired based on the results of an unstructured employment interview. A meta-analysis by Hunter and Hunter (1984) has revealed that the mean true validity of the employment interview is 0.14. Large numbers of applicants (over 200,000) were tested each year. Testing and interviewing costs were set equal to each other, and thus were not considered in utility computations. The job is characterized by unusually low turnover; since no data were available to compute average tenure directly, conservative estimates of five and ten years were used in this study.

Equation (4.1) was used in computing utilities:

$$\Delta U = TN_s(r_1 - r_2)SD_y \phi/p - N_s(C_1 - C_2/p \quad .$$

In the present study, tenure = 5 years or 10 years; number selected = 30 through 130 in intervals of 10; the validity of the test, $r_1 = 0.51$; the validity of the interview, $r_2 = 0.14$; cost of testing per applicant, $C_1 = C_2$. With top-down selection, the term ϕ/p is $0.1758/0.100 = 1.758 = \bar{Z}_x$. When the cutoff score is the mean and applicants above the mean are hired randomly with respect to test score, the effective selection ratio for purposes of computing selection utility is 0.50, and $\phi/p = 0.3989/0.5000 = 0.7978 = \bar{Z}_x$. When the minimum required test score is set at one standard deviation below the mean, the effective selection ratio is 0.84, $\phi/p = 0.2420/0.8413 = 0.2877 = \bar{Z}_x$.

Table 4.11 shows the means, standard deviations, and standard error of estimates. The difference between the two estimates of SD_y is \$1299.97 which is about 34% of \$3800.76 and 25% of \$5100.73. This difference is statistically significant ($p < 0.01$); thus the hypotheses that the dollar value of incumbent employee productivity is normally distributed is not supported by these results. Skewness appears to be negative. The authors propose a number of hypotheses for this finding including ceiling effects on the

Table 4.11. Variable Means, Standard Deviations, and Standard Errors of Estimate

Variable	*N*	*M* ($)	*SD* ($)	*SE* ($)
Performance Level				
Average	114	13530.70	3836.39	359.31
Superior	114	17316.55	5436.98	509.22
Low	114	8346.64	2868.75	268.68
SD_{y1} (S-A)	114	3800.76	2545.77	238.43
SD_{y2} (A-L)	114	5100.73	3813.29	357.15
Number supervised	113	5.49	3.33	—
Experience	113	4.19	1.06	—

Source: Schmidt et al. (1984), p. 494

perceived dollar value of good performance and the relatively greater sensitivity of supervisors to park ranger shortcomings than to superior performance. Table 4.12 examines the intercorrelations among the study variables; although the data fail to support the proposition that employee productivity is really normally distributed, there are some data consistent with the view that supervisors may fail to recognize variations along the entire spectrum of performance. However, the authors note that even if productivity is, in fact, non-normal, the effects of non-normality would be expected to have only minor effects on the accuracy of estimates of the utility of selection (Schmidt et al., 1979; Van Naersson, 1963).

Table 4.13 shows the results of the utility analyses for three modes of test use. For top-down selection, if park rangers are hired by means of a valid test rather than by interviews, the productivity gain is about $1.16 million if they average 5 years on the job, and a gain of about $2.3 million if 10 years on the job.

Table 4.12. Intercorrelations of Study Variables

Variable	2	3	4	5	6	7
1. Average	91*	39*	43*	74*	10	-08
2. Superior		25*	77*	74*	11	-05
3. Low			-05	-29*	01	-01
4. SD_{y1} (S-A)				47*	08	02
5. SD_{y2} (A-L)					12	-07
6. Number supervised						10
7. Experience						

Source: Schmidt et al. (1984), p. 494

Note: Decimals have been eliminated from correlation coefficients.
N = 114, except for Variables 6 and 7, where N = 113.

* $p < .01$.

Table 4.13. Estimated Productivity Increase in Thousands of Dollars From One Year's Substitution of a General Mental Ability Test for the Interview in Selecting U.S. Park Rangers

	Mode of test use and length of tenure (years)					
	Top-down selection		Cutoff score at mean		Cutoff score at -1 *SD*	
Number Selected	5	10	5	10	5	10
30	434	868	197	394	71	142
40	579	1,158	263	526	95	190
50	724	1,447	328	657	118	237
60	868	1,737	394	788	142	284
70	1,013	2,026	460	920	166	332
80	1,158	2,316	525	1,051	190	379
90	1,303	2,606	591	1,182	213	426
100	1,448	2,895	657	1,314	237	474
110	1,,592	3,184	723	1,445	260	521
120	1,737	3,474	788	1,576	284	568
130	1,882	3,764	854	1,708	308	616

Source: Schmidt et al. (1984), p. 495

Table 4.13 also shows that selection utility is substantially reduced when either of the two minimum test score cutoffs is employed. If 80 applicants are hired and average tenure is 10 years, the productivity gain drops from $2.3 million to only $0.38 million for minus one SD below the mean--a gain of only 16% of the top-down gains. The authors raise the question whether employers currently using the low cutoff method are aware of the large price in productivity they are paying.

The authors note, as in the case of previous utility analyses, that the gains will appear to some readers to be overly large and hence implausible. One approach to demonstrating the realistic nature of these figures is to express these values as a percentage of total employee compensation. The value of increased output from improved top-down selection is computed to be 23% of employee compensation costs; for the two minimum test score methods of test use, these figures are 10.4% and 3.8%, respectively.

The selection gains can also be expressed as the percentage increase in total output of new hires. The percentage increase in output due to improved top-down selection of park rangers is computed to be 13%. The percentage increase in total output for the two minimum test score modes of test use are 5.9% and 2.1%, respectively.

The authors conclude that the large dollar gains in output are produced by valid top-down selection resulting from percentage increases in output; by contrast, gains expressed in increase in output may appear modest. Low cutoff methods of test use greatly reduce both dollar gains and percentage increase in output.

6. Productivity Gains in Systems

As described in the previous chapter, Eaton et al. (1985) developed two strategies for estimating the value of performance and for determining *SD*$ by considering the changes in the numbers and performance levels of systems which lead to increased aggregate performance. These strategies attempt to avoid estimation problems in government and in military organizations without private industry counterparts.

Underlying these techniques is the belief that in certain contexts, e.g., tank commanders (TC), supervisors can make more accurate judgments of relative performance than direct estimates of the dollar value of that performance. The underlying payoff scale reflects savings in personnel or equipment costs.

Table 4.14 shows the results for estimates of *SD*$ and examples of utility. The previous chapter described the procedures used to determine *SD*$ for various techniques. As reported earlier, the authors judged that results obtained with global estimations (*SD*$ estimation technique) were unsatisfactory; the 40% of salary proportional rule provided estimates much smaller than other estimates; and the two new strategies, superior equivalents and systems effectiveness, seemed to work well.

Table 4.14. Estimates of *SD*$ and Examples of Utility

	n	*SD*$[a]	*U*$ or utility[a] per tank (N_S = 1)	*U*$ or utility[b] per system (N_S= 2,500)
SD$ Estimation Technique				
Group 1	48	$20,000	$ 4,800	$12,000,000
Group 2	40	$60,000	$14,000	$36,000,000
Superior Equivalents Technique Using Pay and Allowance Estimates of *V*50				
Group 1	52	$26,700	$ 6,400	$16,000,000
Group 2	45	$26,700	$ 6,400	$16,000,000
Using *SD*$ Estimates of *V*50				
Group 1	52	$26,700	$ 6,400	$16,000,000
Group 2	45	$31,100	$ 7,500	$18,700,000
System Effectiveness Technique	-	$60,000	$14,400	$36,000,000
Salary Percentage Technique	-	$12,000	$ 2,900	$7,200,000

Source: Eaton et al. (1985), p. 35

a Rounded to nearest $100.

b Rounded to nearest $100,000.

In computing utilities, a selection ratio of 0.5, $\bar{Z}_x = 0.8$ and $r_{xy} = 0.3$ were used along with *SD*$ estimates for each technique and group. Values of utility (*U*$) for the selection of one tank commander and for the selection of 2500 tank commanders in a

system are given in Table 4.14. Utility per system gains range from $7.2 million for the 40% of salary proportional rule to $36 million for the systems effectiveness technique. The global estimates yielded utilities of $12 million in one group and $36 million in the second group.

As noted earlier, despite its apparent success, the superior equivalents technique may provide underestimates and the systems effectiveness technique may fairly accurately reflect reality. The strengths of the latter technique appear to be based on the availability and interpretability of required data. The authors conclude that the systems effectiveness technique yields estimates that can be adjusted if they appear unreasonable, and such performance and cost figures are subject to open examination and interpretation to a far greater extent than are supervisors' estimates of the dollar worth of various performance levels.

The techniques proposed by Eaton et al. appear to be more useful in situations where individual salary is only a small percentage of the value of performance to the organization or of the equipment operated. Estimates of this technique are appropriate only when the performance of the unit in the system is largely a function of the performance of the individual in the job under investigation. Also as the authors note, the quality of *performance in some situations may not translate into a meaningful, unidimensional,* quantitative scale.

7. The Effects of Variability and Risk on Selection Utility

In the most technically sophisticated selection analysis to date, Rich and Boudreau (1987) conducted the first empirical investigation of parameter variability on utility analysis results. Other utility analysis research, in contrast, derives only point estimates of the expected utility value for selection programs. Utility estimates in most studies are found to be quite large, as a rule, but they fail to reflect the size and shape of the utility distribution and provide little guidance on program riskiness. As the authors note, if two programs offer the same expected return, a rational decisionmaker should prefer the one offering a significantly lower probability of low (or negative) returns and/or a significantly higher probability of very high returns. But utility analysis models generally provide no mechanism for evaluating the relative riskiness and uncertainty associated with different programs.

The present study investigated utility estimate variability for the selection utility of using the Programmer Aptitude Test (PAT) to select computer programmers for

employment at Data General Corporation, a medium-sized firm that manufactures a wide range of computers and peripheral products. Utility calculations were used that incorporated financial/economic factors as well as employee flows over time. The distributions for each utility parameter were empirically estimated; these distribution estimates were combined through a Monte Carlo analysis to yield a distribution of total utility values. Monte Carlo results were compared to three other risk assessment approaches: sensitivity analysis, break-even analysis, and algebraic variance derivation of the distribution.

The authors note that while existing utility formulas do not incorporate utility value variability around the expected value, the notion that utility values represent estimates made under uncertainty has not been completely overlooked (Boudreau, 1983a, 1983b; Cascio & Silbey, 1979; Schmidt et al., 1979). In sensitivity analysis, variability in utility parameters is addressed; each parameter is varied through a range of values, holding other parameter values constant. Combinations of parameter values are examined to determine which one has the greatest impact on the total utility estimate. Such analyses provide no information about the effects of simultaneous changes in more than one utility parameter and also fail to provide the information necessary to assess the probability of observing utility values within a particular range.

The authors describe Boudreau's (1984) extension of sensitivity analysis to determine the lowest value of any individual parameter that would still yield a positive total utility value; these parameter values are termed "break-even values." They represent the values at which the program's benefits are equal to the program's costs. Usually the SD_y parameter is selected for sensitivity analysis because of its variability and questions concerning its validity. Break-even values for SD_y are usually found to be only a very small proportion of the empirically derived expected values (e.g., a $13.12 break-even value compared with a $10,413 expected value for Schmidt et al., 1979, study).

The authors note that break-even analysis compared with expert judgment may simplify measurement and interpretation of utility analyses. It faces limitations, however, that relate to differently shaped utility distributions. For example, if one distribution is more positively skewed than another that has the same expected utility value and similar break-even values for SD_y, then the increased probability of high utility values for the skewed distribution would make it the preferred alternative. Neither traditional utility analysis, sensitivity analysis, break-even analysis, nor algebraic variability derivation (discussed next) would reflect this circumstance.

The authors discuss Alexander and Barrick's (1986) formula for the standard error of utility values associated with a one-cohort selection model, an adaptation of Goodman's (1960) equations for the variance of three or more random variables under conditions of independence. Using data from the Schmidt et al. (1979) study, as well as variance estimates for employee tenure, SD_y, validity, and the number selected, the standard deviations computed by Alexander and Barrick averaged about 50% of the expected utility values. Using break-even analysis and assuming normally distributed utility values, they concluded that in the given context, a selection program had a very high probability of producing benefits exceeding costs.

Algebraic variability derivation is useful for incorporating uncertainty into utility, Rich and Boudreau state, but like the preceding two methods it has limitations for evaluating utility parameters. Limitations of algebraic derivation include the use of formulas which become quite complex when incorporating correlations among different utility parameters, and the assumption of specific distribution characteristics for utility values in making variability estimations. Thus, the algebraically derived utility distribution may be intractable or unrealistic in some decision situations.

Rich and Boudreau state that the fourth risk analysis approach, the Monte Carlo analysis, involves:

> . . . describing each utility model parameter in terms of its expected value and distribution shape. In each trial, a value for each utility parameter is "chosen" from the distribution for that parameter, and the combination of chosen parameter values is used to calculate the total utility value for that trial. Repeated application of this choosing and calculating procedure (using a computer) produces a sample of trials from which the distribution properties of the utility values can be derived. The Monte Carlo procedure addresses the limitations of the other three methods by varying many parameters at once, by incorporating the mathematical interactions among the variables, by providing a mechanism to analyze possible program expansion or abandonment, and by reflecting non-normal distribution assumptions. Of course, Monte Carlo analysis involves more data gathering and computational analysis than sensitivity analysis or break-even analysis (and a different sort of data gathering and computational analysis than algebraic derivation). Therefore, we derived Monte Carlo results and compared them to results from each of the other three risk assessment procedures to empirically examine their relative advantages and disadvantages. (pp. 60-61)

In the present study, entry-level computer programmers were hired by the organization on the basis of an in-plant interview process that consisted of a day-long series of meetings of which only one or two meetings could be characterized as selection

interviews. The proposed change in the process would incorporate an in-plant administration of the PAT, with all other activities remaining unchanged. The PAT data would be used instead of interview data in hiring decisions. Data were available on selection program parameters, employee accession and separation quantities, and some financial/economic investment parameters. These data were incorporated into Boudreau's (1983b) utility model:

$$\Delta U = \sum_{k=1}^{F} \sum_{t=1}^{k} (N_{at} - N_{st}) \left(\frac{1}{1+2}\right)^{k} (r_{PAT} - r_{INT})(\bar{Z}_x)(SD_{sv})$$

$$(1+V)\,(1-TAX_t) - \sum_{\kappa=1}^{F} \left(\frac{1}{1+2}\right)^{k-1} \left(\frac{1}{1+2}\right)$$

$$(C_{PAT} - C_{INT})\,(1 - TAX_t) \qquad (4.3)$$

where:

ΔU = the change in utility associated with replacing interview data with the PAT data, evaluated for F future periods,

F = the number of future periods for which utility is analyzed,

k = the future time period in which utility is evaluated,

t = the future time period in which selectees enter or leave the work force,

N_{a_t} = the number of selectees added to the work force in Future Time Period t,

N_{s_t} = the number of employees separating from the work force in Future Time Period t,

$\bar{Z}_x$ = the average standardized predictor score for the selected group,

SD_{sv} = the standard deviation of dollar-valued service value among the applicant group,

TAX_t = the applicable tax rate for the organization in Future Time Period t,

V = the proportion of service value increased represented by variable costs that change with service value,

SR = the selection ratio,

i = the discount rate,

r_{PAT} = the validity coefficient of the programmer aptitude test (PAT),

r_{INT} = the validity coefficient of the interview (INT),

C_{PAT} = the cost of using the PAT in Future Time Period k, and

C_{INT} = the cost of using the interview (INT) in Future Time period k.

The authors used the tenure distribution of the current computer programmer work force to estimate the likely quantity of separations from each tenure cohort in each future year. The tenure distribution and turnover rates of employee cohorts during the previous six years were used to project the separation/retention and replacement pattern.

To derive the total number of treated-group (newly hired) acquisitions and separations, the authors first estimated the number of separations from the existing work force for Future Years 1 through 6. Then they estimated the number of new hires to fill those vacancies, as well as the number of separations among treated-group employees. Having estimated the separation and acquisition pattern over the program's duration, they were able to estimate the number of treated (newly hired) employees in the work force in each future year as well as the number of years of service provided by those treated employees. The quantity of additions and separations were adjusted to reflect a more realistic amount of service from each acquired or separated cohort, since hiring and leaving do not all occur at the same time of the year. Acquisitions and separations were assumed to be symmetrically distributed around mid-year.

The obtained acquisition/retention pattern was assumed to remain constant in the Monte Carlo analysis; however, variation in the number of employment years was analyzed by varying the number of years of program duration from 4, 5 or 6 years.

The organization had no records from which to estimate selection procedure validity. Hunter and Hunter's (1984) meta-analytic estimate of the true validity of the interview of $r = 0.14$ and a standard deviation of 0.05 was used in the present study, as was Schmidt's et al. (1980) validity estimate of the PAT of $r = 0.73$ and a standard deviation of 0.25. The validity coefficient distributions were assumed to be normal, but because of the high mean and standard deviation of the PAT, a correction was used to re-estimate each coefficient found to be a value greater than one in the risk assessment.

The authors used the ratio of job offers (rather than job openings) to interviews as the selection ratio in order to avoid underestimating the effective selection ratios. The selection ratio over the period from 1980 to 1984 was determined to have a mean of 39.8% and a standard deviation of 4%. For the Monte Carlo analysis, this distribution was assumed to be normal.

To obtain the standard deviation of dollar-valued performance, the authors used a global estimation technique similar to Schmidt et al. (1979). Twenty-nine out of 92 first-line supervisors returned completed questionnaires and were used as subjects.

Table 4.15 lists the questionnaire responses and the estimates derived from them. The mean estimates of the dollar value of the 15th, 50th, and 85th percentiles were \$18,310, \$33,924, and \$50,086, respectively. Averaging the differences between the two extreme percentiles and the 50th percentile produced an estimated SD_{sv} value of \$15,888. This figure represents 60% of the average salary for computer programmers in this organization and, thus, falls within the range of 40-70% suggested by Hunter and Schmidt (1982). As with previous research (e.g., Bobko et al., 1983; Eaton et al., 1985), the present sample of SD_{sv} values was positively skewed, with a median of \$10,000 and a range of \$2,000 to \$60,000.

The authors generated SD_{sv} to reflect two variability possibilities: measurement error or true situational differences. To reflect the latter assumption of SD_{sv} variability (i.e., including both situational variation and sampling error), the authors assumed an SD_{sv} distribution of 29 data points corresponding to the 29 responses of the subjects. Actual SD_{sv} variability is likely to fall between the estimates of sampling error and situational differences, so the two extreme cases should illustrate the maximum impact on utility value variability.

A senior business planning consultant in the organization provided an estimate of the discount rate of 15% per year, excluding the effects of inflation and taxes. The discount rate was assumed constant throughout the Monte Carlo analysis since changes in this parameter are rare. The variable cost proportion that changes with productivity increase, V, was dropped from the analysis because compensation experts indicated that both noncompensation and compensation costs were unlikely to rise if better qualified employees were selected. Although compensation was linked to performance levels, the total amount of compensation costs would remain the same. The organizations' tax rate

Table 4.15. Summary of Survey Responses for Estimating SD_{SV}

Survey number	Percentile Estimates and Differences					
	15th	50th	85th	50th -15th	85th -50th	Average
1	$18,000	$20,000	$22,000	$2,000	$2,000	$2,000
2	21,000	23,000	25,000	2,000	2,000	2,000
3	18,000	23,000	27,000	5,000	4,000	4,500
4	23,000	27,000	33,000	4,000	6,000	5,000
5	20,000	25,000	30,000	5,000	5,000	5,000
6	30,000	35,000	40,000	5,000	5,000	5,000
7	16,000	20,000	27,000	4,000	7,000	5,500
8	17,000	25,000	30,000	8,000	5,000	6,500
9	18,000	27,000	31,000	9,000	4,000	6,500
10	18,000	24,000	32,000	6,000	8,000	7,000
11	18,000	25,000	32,000	7,000	7,000	7,000
12	20,000	30,000	35,000	10,000	5,000	7,500
13	20,000	29,000	36,000	9,000	7,000	8,000
14	20,000	25,300	39,500	5,300	14,200	9,750
15	10,000	20,000	30,000	10,000	10,000	10,000
16	10,000	18,000	30,000	8,000	12,000	10,000
17	30,000	40,000	55,000	10,000	15,000	12,500
18	0	20,000	28,000	20,000	8,000	14,000
19	20,000	30,000	50,000	10,000	20,000	15,000
20	-2,000	19,500	30,000	21,500	10,500	16,000
21	-4,000	18,000	30,000	22,000	12,000	17,000
22	0	25,000	35,000	25,000	10,000	17,500
23	0	20,000	50,000	20,000	30,000	25,000
24	25,000	50,000	75,000	25,000	25,000	25,000
25	75,000	100,000	130,000	25,000	30,000	27,500
26	30,000	70,000	100,000	40,000	30,000	35,000
27	60,000	85,000	150,000	25,000	65,000	45,000
28	0	50,000	100,000	50,000	50,000	50,000
29	0	60,000	120,000	60,000	60,000	60,000
Average	$18,310	$33,924	$50,086	$15,614	$16,162	$15,888
SD	$16,797	$20,407	$34,469	$14,135	$16,558	$14,617

Source: Rich and Boudreau (1986), p. 68

over the previous three years had been 38%, 39% and 40%, with a mean value of 39%. The distribution of tax rates was assumed to be a uniform distribution with each of the three values being equally likely. The 1000 generated trials produced a standard deviation of 0.006.

The cost of the interview alone was estimated to average $634 per candidate, while the cost that included the PAT was estimated to be $644. Per-candidate cost variability for the interview was generated by constructing a uniform distribution of interview cost levels between $534 and $734, in increments of $0.01.

Rich and Boudreau describe the Monte Carlo procedure:

> Utility parameter values and total utility estimates were generated by computer program written with the Statistical Package for Social Sciences (SPSS), for AOS/VS, Version M, Release 9.0 (Nie, Hull, Jenkins, Steinbrenner, & Bent, 1975). Each generated trial produced one utility estimate. Within each trial, values for the different utility parameters were generated from populations with the distribution characteristics discussed above. Then, these utility parameter values were combined to produce a total utility estimate reflecting the discounted, after-tax benefits less costs of both selection programs, and the differences between them. The formula used by the Monte Carlo analysis to compute the total utility difference is [a modification of Equation 4.3]. . . .
>
> For each trial, the computer program first draws a value indicating the duration of the program (i.e., 4, 5, or 6 years). This value determines which one of three cohort-by-future year tables (similar to the lower portion of Table 3) will apply to that trial. The values for EY_{ct} come from this table. Next the computer program generates a selection ratio (i.e., SR_c for each cohort to be acquired and (using table values from Naylor & Shine, 1965) converts that selection ratio into a standardized predictor score (i.e., $\bar{Z}_x$ for each cohort. Then, the program generates values for those parameters assumed constant across all cohorts and time periods in each trial (i.e., SD_{sv}, r_{PAT}, r_{INT}, TAX, C_{PAT}, and C_{INT}). For each cohort-year combination, the program computes the one-year incremental value and selection cost for that cohort using Equation 3. Finally, using the discount rate (i = 0.15) as shown in Equation 3, the program sums across time periods and cohorts to obtain the overall utility value. This procedure was employed to generate 1,000 trials. (pp. 71-72)

Rich and Boudreau present calculations shown in Table 4.16 of the expected utility values for the PAT, the interview, and the differences between them. The lower portion of Table 4.16 indicates that the PAT produces a utility (i.e., $3.20 million) over seven times as large as the interview (i.e., $.43 million), a difference of $2.77 million. Due to the flow of employment years, the largest benefits from each predictor are realized in years 3 through 5. These utility values indicate a substantial financial advantage to investing in the PAT.

Table 4.16. Expected Utility Value Calculations

	Constant factors across both programs						Factors varied across programs			
t	EY_t	DF	SD_{sv}	$\bar{Z}_x$	$1-TAX$	SR	r_{INT}	r_{PAT}	C_{INT}	C_{PAT}
1	25	0.87	$15,888	0.97	0.61	.398	0.14	0.73	$634	$644
2	79	0.76	15,888	0.97	0.61	.398	0.14	0.73	$634	$644
3	129	0.66	15,888	0.97	0.61	.398	0.14	0.73	$634	$644
4	167	0.57	15,888	0.97	0.61	.398	0.14	0.73	$634	$644
5	190	0.50	15,888	0.97	0.61	.398	0.14	0.73	$634	$644
6	168	0.43	15,888	0.97	0.61	.398	0.14	0.73	0	0
7	110	0.38	15,888	0.97	0.61	.398	0.14	0.73	0	0
8	60	0.33	15,888	0.97	0.61	.398	0.14	0.73	0	0
9	26	0.28	15,888	0.97	0.61	.398	0.14	0.73	0	0
10	8	0.25	15,888	0.97	0.61	.398	0.14	0.73	0	0
11	1	0.21	15,888	0.97	0.61	.398	0.14	0.73	0	0

	Total discounted, after-tax utility vales (millions)			
t	ΔU_{PAT}	ΔU_{INT}	$\Delta U_{PAT} - \Delta U_{INT}$	
1	$0.10	-$0.02	$0.12	
2	0.36	0.03	0.33	
3	0.54	0.07	0.47	
4	0.61	0.08	0.53	
5	0.61	0.09	0.52	
6	0.50	0.10	0.40	
7	0.28	0.05	0.23	
8	0.13	0.03	0.11	
9	0.05	0.01	0.04	
10	0.01	0.00	0.01	
11	0.00	0.00	0.00	
	$3.20	$0.43	$2.77	Total Program Utility

Source: Rich and Boudreau (1986), p. 72

Note: DF equals $\left(\frac{1}{1+2}\right)^t$.

Table 4.17 summarizes the results that assume variability is due to sampling error (i.e., the SD_{sv} values come from a normal distribution with a mean of \$15,888 and standard deviation of \$2,761). The results also assume variability that includes situational differences observed by respondents (i.e., SD_{sv} values with a positive skew produce a mean of \$15,888, a median of \$10,000 and a range of \$2,000 to \$60,000).

Table 4-17. Summary Descriptive Statistics Derived from the Monte Carlo Analyses

	Interview	PAT	PAT - INT
Analysis assuming SD_{SV} variability is due only to sampling error			
Average ΔU	\$0.42	\$3.16	\$2.74
Median ΔU	0.49	3.12	2.73
Lowest ΔU	−0.38	−1.70	−2.21
Highest ΔU	1.97	7.76	6.28
$SD\,\Delta U$	0.28	1.35	1.31
Analysis assuming SD_{SV} includes situational differences			
Average ΔU	0.41	3.18	2.77
Median ΔU	0.38	1.88	1.70
Lowest ΔU	−0.66	−0.31	−0.84
Highest ΔU	4.16	20.74	18.64
$SD\,\Delta U$	0.68	3.56	3.09
Expected utility values calculated with the traditional utility formula			
Expected ΔU	0.43	3.20	2.77

Source: Rich and Boudreau (1986), p. 74

Note. All values expressed in millions

In both analyses the average utility values are quite similar to the values calculated in Table 4.16. The major difference between the two analyses is in the shape and size of the utility distribution. Both analyses seem to suggest a low probability that the PAT will produce negative utility or that it will fail to produce greater utility than the interview. They do, however, suggest very different outcomes in the high end of the utility distribution.

Comparing Monte Carlo results to sensitivity analysis, Rich and Boudreau conclude that a risk-aversive decisionmaker might incorrectly decide to "play it safe" and not risk the new selection procedure. Also, the sensitivity analysis provides no information regarding the probabilities of expected values.

Using Boudreau's (1984) methodology for calculating the minimum value of a particular utility parameter necessary to produce total utility values greater than zero, the authors found a break-even yearly SD_{sv} for the interview of $5,451.86; $1,062.03 for the PAT; and $20.40 for the utility difference. Thus, the authors note, for every SD_{sv} value above $20.40, the PAT will produce greater utility than the interview, and PAT utility will exceed zero if SD_{sv} exceeds $1,062.03. The interview utility does not exceed zero unless SD_{sv} exceeds $5,451.86. However, this type of analysis assumes the other utility parameters remain fixed at their expected levels, and it provides no information on the probability of attaining negative utility values with the proposed selection program.

In comparing Monte Carlo results to algebraic variability estimates, the authors modified Alexander and Barrick (1986) one-cohort utility model to reflect the implications of subsequent flows of employees through the work force and repeated applications of a selection program (Boudreau, 1983b). The algebraic derivation, like the other three risk-estimates methods, suggests that the PAT is a relatively safe investment (i.e., the 95% confidence values all exceed zero). Compared to the Monte Carlo analysis, however, it provides much less detailed information and requires assumptions regarding correlations among cohort-year utility values.

The authors conclude:

> The very modest additional cost of the PAT and its substantially greater validity led all of the variance estimation methods to suggest the same decision--implement the PAT. However, the implied probability of extreme utility values differs greatly with assumptions about variability in SD_{sv}. Only the Monte Carlo analysis captures this information. Moreover, in situations with higher program costs, more similar program effects, or competing investments the Monte Carlo information, might well alter utility analysis conclusions. Future enhancements to the risk assessment models might also incorporate internal and external employee movement patterns, or different sources of predictor validity distribution estimates. We hope this initial application will motivate such future investigations. (p. 82)

C. UTILITY ANALYSIS RESULTS AS DECISION AIDS

Since the earliest utility analyses using rational estimates in the late 1970s, the value of selection programs has been found to be quite large--in the millions of dollars per year.

One of the major objectives of these early utility analyses was to make just that point through demonstrations of productivity gains in understandable quantitative terms.

Some investigators became concerned that estimates of dollar-valued performance due to selection procedures may be seen as overly large or even implausible. Various alternative measures were proposed to demonstrate the realistic nature of productivity gains by expressing gains in percentage of increase in output or of employee compensation rather than in dollars.

Later in the 1980s, concern shifted to consideration of the psychometric properties of the payoff scale, principally to its variability and accuracy or validity. At least six major approaches were proposed to measure the payoff scale, SD_y, to improve its reliability, understandability and credibility. Despite a fair number of comparative analyses of the various approaches, controversy still surrounds the measurement of SD_y and no standard approach is universally accepted, although the global estimation approach is used most frequently.

In about the same time frame, a number of modifications to the general utility analysis equation were proposed that would better reflect economic factors including variable costs, tax liabilities, and discount rates. Consideration of the combined effects of these economic variables would considerably reduce reported estimates of productivity gains of most studies. On the other hand, consideration of employee flows or the program effects on more than a single group of applicants would considerably increase reported estimates of productivity gains of most studies.

Recently, a new concern in utility analysis was signaled by Rich and Boudreau (1987), in their suggestion that decisionmakers may be aided by guidance relating to a program's riskiness. In an empirical test described earlier in the chapter, a Monte Carlo analysis proved superior to other risk approaches, including sensitivity analysis, break-even analysis, and algebraic variance derivation. Even though results indicated the same decision would have been made based on utility elements of any one of the four analyses, the additional effort required by the Monte Carlo analysis always led to a correct analysis whereas any one of the other methods may not have done so.

Conventional belief asserts that selection procedure decisions are very unlikely to be affected by the type of SD_y estimate used or by the size and shape of the utility distribution. This assertion is likely to be true when the selection procedures refer to comparisons between a test, an interview, an assessment center or between top-down selection or

selection using a low cutoff score. In such types of comparisons, differences in utility estimates usually are very large. However, there are real-world decision contexts in which risk analysis in utility estimates may be quite important in decisionmaking. In the military, for example, in selection and job classification, setting an ability standard within a percentile point has important productivity consequences because of overall productivity gains and costs associated with treating a large number of individuals. Employing actual variability in individual utility parameters in a Monte Carlo approach in this context would assess the riskiness in setting the standard and could change the decision or confidence in the decision. Research aimed at gaining better understanding of the information used in the decision process and in assisting the process continues to be much needed.

CHAPTER 5. NEW USES AND EXTENSIONS OF THE BASIC UTILITY MODEL

This chapter addresses recent changes in the basic utility model by:

- Reviewing an alternative expression of the basic utility model that makes it applicable to all personnel interventions aimed at improving job performance;
- Describing break-even analysis, an approach designed to simplify information required for decisions; and
- Examining extensions of the basic utility model that incorporate economic factors and the effects of employee flows.

At this point in our review and analysis of selection utility, the knowledgeable reader will readily recognize our indebtedness to the original contributions of several preeminent investigators. It is fitting, then, to acknowledge them in this chapter, which comments on current formulations of selection utility models.

Hubert Brogden provided in 1946 the most widely accepted decision-theoretic interpretation of the validity coefficient. This pioneering derivation was followed shortly by his classic formulation of the basic selection utility model in dollar-valued terms. In the 1950s, Brogden (along with Paul Horst) developed much of the theory and methodologies used in determining classification decisions and in maximizing classification efficiency described in subsequent chapters.

Frank Schmidt and John Hunter are widely credited for fostering the current high interest in selection utility analysis. Despite the availability of Brogden's model for thirty years, it received little attention. Schmidt and Hunter attributed the lack of use of the model to misconceptions concerning statistical assumptions, belief that validation studies considered necessary for each application were costly, and the belief that difficult cost accounting procedures were necessary for estimating dollar-valued performance. Schmidt and Hunter presented findings that dissipated or refuted the first two concerns and developed a practical means of estimating performance in dollar terms that facilitated use of utility analysis. Through a series of highly innovative studies, they demonstrated the great economic value of valid selection procedures on work force productivity. They also

developed a version of the basic utility model that applies to all personnel interventions designed to improve job performance.

John Boudreau is widely acknowledged to have introduced greater realism in utility models by incorporating considerations of financial and/or economic factors and the effects of employee flows. Such extensions provide more precision in estimating utilities. Boudreau stressed procedures that aid decisionmaking rather than those that simply demonstrate utility value. He introduced risk analysis techniques to further enhance credibility of results used in decisionmaking. Boudreau's research has contributed to greater awareness of utility model assumptions and limitations as applied in specific decision contexts.

Wayne Cascio is widely recognized for his accessible writings on utility concepts and for his demonstrations of practical applications of costing human resource programs in business. He has contributed an alternative technique to Schmidt and Hunter's for estimating dollar-valued performance and has conducted important empirical evaluations of parameter estimation variability and accuracy. He is an early and forceful advocate of an interdisciplinary approach in utility analysis and of formulating models incorporating economic theories of organizations as a means of enhancing credibility of "bottom-line" assessments of personnel programs.

A. ALTERNATIVE APPLICATIONS BASED ON THE GENERAL UTILITY MODELS

Brogden's (1949) historic equation is the building block for all recent extensions on elaborations of utility analysis. The following formulation has minor notational variations from equations presented in Chapter 2.

$$\Delta U = (N)(T)(r_{x,y})(SD_y)(\bar{Z}_x) - C \qquad (5.1)$$

where:

ΔU = the increase in average dollar-value payoff that results from selecting N employees using a test or procedure (x) instead of selecting randomly,

N = the number of employees selected,

T = the expected tenure of the selected group,

$r_{x,y}$ = the correlation coefficient (among prescreened applicants or incumbents) between predictor score (x) and dollar-value payoff (y).

SD_y = the standard deviation of dollar-value payoff in the group of prescreened applicants,

$\bar{Z}_x$ = the average standard predictor score of the selected group, and

C = the total selection cost for all applicants.

1. Generalization of the Basic Utility Model to Other Intervention Programs

In recent years, it was suggested that Brogden's utility model is applicable to any type of personnel program designed to increase the job performance treated by the program (i.e., the intervention offered to a group in order to improve performance). Schmidt et al. (1982) showed that the product of $r_{x,y}$ and $\bar{Z}_x$ in Equation (5-1) may be replaced by the "true differences in job performance" (i.e., correction for criterion unreliability) in standard deviation units between the treated and untreated groups. The resulting utility formula is given below:

$$\Delta U = (N)(T)(d_t)(SD_y) - C \qquad (5.2)$$

where:

ΔU = the increase in utility resulting from the program,

N = the number treated,

T = the expected duration of benefits in the treated group,

d_t = the true difference in job performance between the treated and untreated groups in standard deviation units,

SD_y = the standard deviation of dollar-valued job performance among the incumbent employees, and

C = the cost of treating N employees.

The purpose of the Schmidt et al. (1982) study was to illustrate how the general utility model used to evaluate selection procedures could be adapted to the evaluation of a hypothetical computer programmer training course in dollar terms. All the utility parameters needed for Equation (5-2) were available from an earlier study (Schmidt et al., 1979), except for d_t,C, and T.

A value of $d = 0.50$ in standard score units was assumed as the difference in performance ratings between the trained and untrained groups. Correcting for unreliability in the ratings (King, Hunter & Schmidt, 1980) produced a true difference, $d_t = 0.65$.

The authors noted that the value of d_t can be estimated from other studies that provide only F values by converting the t statistic into a point-biserial correlation and then converting the value of r into d_t by the use of several simple formulas:

$$r = \frac{t}{\sqrt{t^2 + [N_t - 2]}}$$

where N_t is the total number of persons in the study (that is, the sum of the experimental and control groups). The value for r can be converted to d using the following formula:

$$d = \frac{1}{\sqrt{pq}} \qquad \sqrt{\frac{N_t - 2}{N_t}} \cdot \qquad \frac{r}{\sqrt{1 - r^2}}$$

where p and q are the proportions of the total group in the trained and untrained groups, respectively.

In evaluating organization interventions, the relevant group was incumbent employees--those receiving the intervention (e.g., training). To use the (unrestricted) applicant group, as is the practice in evaluating selection programs, would overestimate SD_y. The appropriate value of \$10,413 per year for SD_y was obtained from the Schmidt et al. (1979) study.

In evaluating cost of training, the authors included only direct training costs because they assumed training sessions are not held during working hours; if training were held during working hours, that additional cost would have to be added to the other training costs. The cost figure of \$500 per trainee assumes that the programmers take the training course only once.

In determining the value of T, the duration in years of the training effect, the authors noted that intervention effects decline gradually rather than disappear abruptly. They estimated that programmer training might decline to zero over a period of four years. Taking this into account, they speculated that the best estimate of T might be the duration of the period of decline divided by 2, leading to $T = 2$. Inserting all parameter estimates into Equation (5.2) produces:

$$\Delta U = (100)(2)(.65)(\$10{,}413) - 100(\$500)$$

$$\Delta U = \$1{,}303{,}690 \ .$$

The authors also computed variations of Equation (5.2) permitting evaluation of continuing intervention (readministering a program periodically) and comparisons among a number of different interventions (rather than forcing a choice between no intervention and an intervention).

How realistic is Schmidt et al.'s example that assumed a $d_t = 0.65$ [corresponding to a (corrected for reliability) point-biserial of 0.31 between the trained versus non-trained dichotomy and job performance]? The authors have cited a number of reviews and studies examining effect sizes for different types of interventions that show d_t ranging from 1.00 to 0.10 standard score units. Based on these findings, the authors believe that the d_t value of 0.65 assumed in their study was not unreasonable.

In a recent meta-analysis, Burke and Day (1986) examined the effect size estimates of five managerial behavioral modeling techniques against subjective behavioral criteria. They found an average effect size estimate of 0.70 (SD = 0.52). After correction for statistical artifacts, the average effect size was estimated to be 0.78 (SD = 0.00).

The optimal strategy for combining selection and performance enhancement effects (which in some situations may be interactive), Schmidt et al. (1982) conclude, is always first to maximize the effectiveness of selection and then to apply intervention programs that are most effective in increasing further the performance of those selected (i.e., intervention programs that work best with high ability incumbents).

Landy et al. (1982) also illustrated the generalized utility model by demonstrating the utility of a hypothetical performance evaluation and feedback intervention program for managers. The authors assumed values for insertion into Equation (5.2) were $N = 500$, $T = 1$ year, $SD_y = \$20{,}000$, and $C = \$700$ per employee. The value for d_t was not given, but was based on a conservative estimate of "validity" of 0.30 for the intervention by use of the equation proposed by Schmidt et al. for transforming an r into d. (By substitution, we find $d_t = 0.565$.) The authors used the term "validity" because they assumed that alternative strategies for evaluation and feedback are differentially "valid." The estimate of SD_y was a global estimate determined informally on the basis of conversations with executives. It was judged a rather primitive estimate because the executives suggested the same estimate figure for all first-level middle managers regardless of responsibility. The cost of training supervisors on the intervention technique was considered by the authors as an overestimate since typically several managers were evaluated by a single supervisor.

The utility or productivity gain was computed to be $5.3 million a year--a reasonable investment by any standard, the authors assert. Landy et al. conclude that by the use of Equations (5.1) and (5.2):

> It appears that it is possible to view the entire system by which organizations select, train, place, and motivate employees from a utility perspective. This is so because the object of each of these strategies is to increase the mean performance of a potential work force. If we know that the standard deviation of performance in dollars is of a given value, and if we know the value of average performance in dollars for "treatment" and control groups, then it is possible to use utility calculations to determine how much it would cost to move mean performance up one standard deviation. This argument simply recasts a correlation problem as a regression one, i.e., costs are analogous to regression weights. Utility might be thought of as the dependent variable and each of the potential strategies for increasing that utility as independent variables. (p. 32)

Mathieu and Leonard (1987) conducted an operational empirical evaluation of a training program in supervisory skills on the performance ratings of bank supervisors. They used an expanded version of Equation (5.2) to take into account the influence of turnover, diminishing effects of training, and estimated costs over one-year periods. The equation was also adjusted to account for economic considerations; both modifications followed Boudreau's formulations, discussed in detail in a later section of this chapter. The complete equation used was:

$$\Delta Uk = \sum_{g=1}^{G_k} \frac{1}{[1+i]^k} [N_{gk} SD_y (1 + V) (1 - TAX_k) d_{tgk}] - \left[Ck \frac{1}{[1+i]^k} (1 - TAX_k) \right]$$

where:

k = the number of years over which utility estimates are calculated,

ΔU_k = the marginal utility gained in year k,

G_k = the total number of groups trained through year k,

N_{gk} = the number of trainees in group g in year k adjusted for turnover,

SD_y = the standard deviation of performance in dollar units,

d_{tgk} = the effect size estimate for training group g in year k,

C_k = the costs incurred in year k,

i = the discount rate,

V = the variable costs, and

TAX_k = the organizational tax rate for year k.

The authors noted several features of their modified equation, including: a more complete and precise definition of utility than existed for earlier models; an equation feature allowing both the number of treated employees in a work force and the effect size to vary according to the time of treatment; the replacement of the selection effects with d_t, the appropriate size estimate for a training intervention; and the use of discounted costs associated with training in the same period in which benefits would begin to accrue.

The subjects for the study were 65 employees of a bank that had completed a training program in supervisory skills in the previous year. Because individuals had not been randomly assigned for training, a control group of the same size was matched on all performance-relevant variables available. Comparisons made between the trained and control groups on the matching variables revealed no significant differences, including performance appraisals completed prior to the start of the training date for both groups.

The authors obtained dollar-valued SD_y estimates from supervisors using the global estimation method (Schmidt et al., 1979) separately by job classes (head teller, operating manager and branch manager). Supervisors were also asked to evaluate subordinates on an 18-item performance rating scale. A composite score was used as the measure of overall job performance.

One-time training costs were estimated to be $12,800. Variable costs for the three-job classes ranged from $367 to $601. Turnover for the three-job classes ranged from 10.5% to 16.9%. The discount rate was determined to be 15% and the tax rate, 46%. Salary was considered to be the only variable cost associated with improvement in training; therefore utility gains from training were reduced by a percentage calculated by dividing each job's variable costs by its SD_y value.

Mathieu and Leonard estimated d_t as 0.3146 based on a hierarchical multiple regression to determine the influence of training performance, and an equation transforming the partial correlation representing the independent effect of training into d. The authors noted that the estimated value was in fact, not d_t but d, since they had not corrected the measure for attenuation. Thus, the difference score represents a conservative estimate of the true effect of training on performance.

The SD_y estimates were found to vary widely within each job class. After the distribution was trimmed for outliers, no significant difference emerged from *t*-tests. The final averaged SD_y for each class used in utility computations ranged from \$2,369 to \$10,064. The "raw" estimates of utility (unadjusted for economic considerations) of training 65 employees was \$78,493 for the first year alone, the benefits continuing to rise to \$421,427 by year 5 and \$750,883 by year 20. Reducing utilities for economic factors was seen to result in "drastic" reduction (e.g., to \$194,885 by year 20).

The authors suggested that perhaps the most tenuous assumption they made was that the effects of training on performance (d_t) remain constant over time; the alternative argument is that the effects of training on performance dissipate over time. The authors therefore computed both raw and adjusted utility estimates, assuming a 25% reduction in d_t each year. Results showed that in the conservative (adjusted) case of training declines, utility would fall to \$105,852 by year 20.

The results presented thus far represent only the benefit from training 65 supervisors. Thus these findings underestimate the true value of the program since the analysis did not include the influence of training additional groups. For example, a decisionmaker might ask: What would be the summed overall utility of the program if training were conducted with additional groups for five years?

The turnover rates used in the analysis showed at least 15 openings occurring each year in each job class; on the basis of 45 employees, the estimated adjusted utility of the program in five years would be \$219,577. The estimate for the tenth year of the utility of training 225 employees in the first five years would be over \$364,300, after adjustments for economic considerations. Utility estimates would, of course, be lower under the assumption that the effectiveness of training diminishes 25% each year (e.g., \$213,334 by year ten).

The authors concluded that their findings were compelling not only in terms of dollars, but also from the standpoint of information provided for managerial decision-making.

Mathieu and Leonard's (1987) study is an important, realistic application for several reasons: it was the first empirical utility analysis of the organizational benefits of training employing Equation (5.2), the revised version of the basic model; it employed an expanded version of the utility model incorporating economic factors and employee flows; it demonstrated the problems involved in a quasi-experimental design to assess the effect of

training and estimate the diminishing return from the effect; and it employed risk analysis. Its results provide credible estimates clearly showing the cost-effectiveness of the training program.

2. Break-Even Analysis: Simplifying Information for Decisions

Boudreau (1984) suggested that in some decision contexts a choice can be facilitated between one selection procedure and another by determining the lowest value of any given parameter that would still ensure that the total utility of the preferred procedure is at least equivalent to the other procedure. Obtaining added information would not only affect no change in the decision, but would likely incur added costs. This reasoning, Boudreau notes, is common in microeconomic theory.

The value of additional information in choosing among alternatives in a financial management context was investigated by Bierman, Bonini and Hausman (1981). Bierman et al. proposed that identifying the value of additional information involves specifying a "break-even point." Choosing an alternative above the break-even point results in positive payoff; choosing one below that point results in negative payoff; and choosing one at the break-even point results in zero payoff.

Boudreau applied the break-even approach to selection utility, pointing out that instead of estimating the level of expected utility for each alternative, identifying the break-even values critical to making a decision would be simpler. The value of additional information (precision) for some utility parameters in a given decision situation may be quite low because added precision measurement would not alter the choice among alternatives. In such situations, Boudreau asserts, utility models may be more practical decision tools than previously thought. Traditional emphasis in utility analysis is placed on demonstrating estimated utility values; the emphasis in break-even analysis is on making decisions. Break-even analysis can specify the minimum parameter values required for decisions, often without the necessity of estimating utility values at all.

Boudreau (1984) provided a number of examples of break-even analysis for both simplifying the number of choices or choosing among alternatives that were considered in Schmidt et al.'s (1979) study, described in the previous chapter. Sometimes the less attractive alternative is "event dominated" by the more attractive one. In evaluating the utility of the Programmer Aptitude Test (PAT), the validity coefficient and cost emerged as the only two relevant decisionmaking variables. As Equation (5.1) shows, validity is proportional to dollar benefits. Given a situation in which one alternative (the PAT) has

higher validity under all conditions, and an equal or lower cost than a second alternative (e.g., the interviewer), the first alternative always will dominate. In choosing between the two, one alternative could be eliminated without any other information, including that from obtaining estimates of SD_y.

Because the PAT is more valid but more costly than random selection, two decision options pertain: random selection or selection by means of the PAT. Break-even analysis still can be used to simplify the decision. As described earlier, the SD_y determined in the Schmidt et al. (1979) study was $10,413. Boudreau assumes, for illustrative purposes, that SD_y is unknown in order to show how its critical values can be established. Substituting the known parameters into Equation (5.1), the utility of the PAT is:

$$\Delta U_{PAT} = (618)\,(9.69)\,(.80)\,(.76)\,(SD_y) - (\$10)\,(618/.50)$$

where $N = 618$; $T = 9.69$ years; $\overline{Z} = 0.80$, when SR = 0.50; $r_{xy} = 0.76$; and the cost of testing, $C = \$10$ per applicant. This equation simplifies to

$$\Delta U_{PAT} = 3{,}641\,(SD_y) - \$12{,}360 \quad .$$

In the case of random selection, the decisionmaker can either choose to adopt the PAT or not adopt it (i.e., use random selection). Boudreau asks what value is needed for the unknown parameter, SD_y, in order to produce positive ΔU. This is answered by substituting a zero for ΔU in the above equation, producing

$$0 = 3{,}641\,(SD_y) - \$12{,}360;\ SD_y = \$3.39 \quad .$$

For this decision, then, the critical question is whether or not SD_y exceeds the break-even value of $3.39 a year.

Schmidt et al. showed utility analysis results for a range of selection ratios and examined the sensitivity of utility values to changes in the selection ratio. Boudreau computed the break-even values of SD_y for various selection ratios using Schmidt et al.'s data. Results show that the SD_y for values required to produce positive utilities are not very large relative to the $10,413 SD_y value reported in the original study. The highest (most conservative) break-even SD_y value, when SR = 0.05, is $13.12 per year. The lowest (least conservative) break-even SD_y value, when SR = 0.50, is $3.39. Because the most conservative break-even value of $13.12 is so low (7.78 standard deviations below the mean), it was probably not necessary, the author asserts, to obtain more precise

estimates of SD_y since the aim of the break-even analysis was to produce only the basic information needed to make the decision.

Later studies reported similarly low break-even points. For example, Burke and Frederick (1986) found the break-even value to be 34% of the average of seven SD_y estimates; Mathieu and Leonard (1987) found break-even values ranging from 13% to 50% of SD_y estimates based on very conservative parameter values; and Rich and Boudreau (1987) found in their risk analysis study that the break-even value for the PAT was about 6% of the mean SD_y estimate. If one were to compute break-even points for earlier published studies, it appears nearly certain that decisions whether or not to adopt the programs would have been unaffected. This is so because reported productivity gains had been uniformly high, even after measurement errors affecting the magnitude of SD_y estimates had been taken into account.

Many decision situations involve a number of mutually exclusive "undominated" alternatives to random selection procedures. The same type of break-even analysis applies in deciding among multiple alternatives. Separate break-even utility equations would be computed for each alternative, permitting comparisons of break-even parameter values among the alternatives.

Boudreau (1984) suggested a number of advantages of break-even analysis: the parsimonious use of information; the relative ease in making threshold (break-even) judgments compared to estimating actual SD_y values higher than a threshold value, even though judges are unlikely to agree on the exact point estimate for the SD_y parameter; and greater understanding of how even small SD_y values can produce sizeable utility gains.

In short, although break-even analysis is a simple approach that is aimed at aiding decision making, it is not, as mentioned in the previous chapter, the best approach for dealing with risk and uncertainty (Rich & Boudreau, 1986).

Employing break-even analysis without estimating utility gains, a procedure suggested by Boudreau for some decision situations is of greater concern. If this break-even analysis "advantage" is actually used alone, then organizational opportunity costs are explicitly ignored. If choosing the preferred alternative yields small productivity gains (e.g., replacing an existing selection procedure by a new one based on utilities estimates), the organization may decide to invest its limited resources in some entirely different program that may result in greater net gains than the proposed selection alternative.

For example, choosing a higher enlistment standard than the existing one for determining entry eligibility into the Army may result in a net benefit of $300 million, even after taking into account the $500 million in additional recruiting costs required to attract "higher quality" individuals. (These are not unrealistic figures as will be empirically demonstrated in a later chapter.) However, manpower policymakers may be persuaded that it would be better to invest the additional $500 million needed for recruiting costs on purchasing improved helicopter flight simulators for training pilots. Acquiring such simulators, they might reason, would reduce high training attrition, save lives and costly equipment and improve the operational effectiveness of pilots--benefits estimated to achieve greater productivity gains than those achieved by recruiting high quality individuals.

Where competing investments for global resource allocation are involved, the very modest additional cost of obtaining more information on utility values and some type of risk assessment appears clearly warranted.

Boudreau acknowledges, of course, that break-even analysis is not a substitute for utility values. He also suggests that additional measurement precision prediction is warranted in decisions where break-even values fall very close to the best prior parameter estimate, where that estimate is extremely uncertain, or where the loss function is very steeply sloped.

B. EXTENSIONS OF THE BASIC UTILITY MODEL

This section examines extensions of the basic utility model that incorporate financial factors and the effects of employee flows.

1. Financial Accounting Considerations in Estimating Utility

Boudreau (1983a) extends utility formulas by incorporating three financial and/or economic considerations: variable costs, taxes, and discounting. Utilities that include economic factors are more directly comparable with utilities of other management functions. Boudreau notes that previous utility studies defined the payoff function, SD_y, as the "value of sales" (Cascio & Sibley, 1979); or the "value of products and services" (Schmidt et al., 1979); or the value of "output as sold" or "what the employee charges the customer" (Hunter & Schmidt, 1982). With regard to variable costs, Boudreau states that increases in the value of productivity or "sales value" of the productivity will misstate the institutional benefit of productivity increases when variable costs rise or fall with productivity increases (e.g., incentive or commission-based pay, benefits, variable raw

material costs, and variable production overhead). Utility estimates based on such a "deficient" payoff definition may produce large biases when compared to payoff estimates for other investments.

Boudreau states that when a selection decision results in increases in productivity, the decision may also increase or decrease costs associated with productivity; variable costs should be subtracted from or added to the increased value attributable to increased productivity. For example, better-selected sales people may sell more, but not all of the increase in sales revenue accrues to the firm, since the higher-productivity salespeople often receive greater pay, bonuses or commissions. Thus the benefits of increase in productivity are less than the increases in sales revenue. Conversely, better-selected employees may also incur fewer costs, (e.g., operators that reduce wastage would augment the sales value of productivity).

Boudreau modified Equation (5.1) to include factors that account for sales values, sv_j (e.g., sales revenue), service costs, sc_j (e.g., wages, materials), and the net benefits, nb, the difference between sv_j and sc_j (i.e., $nb_j = sv_j - sc_j$). The three terms are thus random variables over the population of preselected applicants or incumbents. He argues that "net benefits" is a more logically correct definition of the payoff value in Equation (5.1) than is sv or sc. Equation (5.1) then becomes:

$$\Delta U = (N)(T)(r_{x,nb})(\bar{Z}_x)(SD_{nb}) - C \qquad (5.3)$$

Moreover, it can be shown that Equation (5-3) is equivalent to

$$\Delta U = (N)(T)[(r_{x,sv})(\bar{Z}_x)(SD_{sv}) - (r_{x,sc})(\bar{Z}_x)(SD_{sc})] - C \quad . \qquad (5.4)$$

Boudreau notes that Equation (5-4) recognizes that a selection device, x, in correlating with sales volume, may also correlate with service costs. In many situations, it may be simpler to treat service costs as perfectly correlated with sales value. In this case, $r_{x,sv}$ is equal to the absolute value of $r_{x,sc}$ treated as a proportion of sv (e.g., where commissions equal a percent of sales revenue or variable material costs comprise a percent of selling price); then Equation (5-4) becomes:

$$\Delta U = (N)(T)(r_{x,sv})(\bar{Z}_x)(SD_{sv})(1 + V) - C \qquad (5\text{-}5)$$

where V equals the proportion of sv represented by sc (i.e., $sc/sv = V$). This parameter (V) will be negative when a higher proportion of costs varies positively with sales value, and positive when a higher proportion of costs varies negatively with sales values.

Boudreau further suggests that fixed costs are irrelevant because the utility of a selection device will not change fixed costs, (i.e., fixed cost variability is zero). Compensation costs and variable costs other than compensation, on the other hand, may vary positively or zero with productivity. When salaries and benefits vary positively with sales value (sv), then the standard deviation of sales value SD_{sv} will overestimate the standard deviation of net benefits (SD_{nb}). The major non-compensation cost which varies with sales value of productivity (sv) is probably raw materials. When such costs vary positively with sv, it would further reduce V; when non-compensation costs vary negatively (e.g., wastage), it would increase V and thus sales value.

Combining all arguments relating to the effects of variable cost, Boudreau suggests that assuming a range of V from – 0.50 to + 0.33 would not be unreasonable. This range implies an adjustment [1 + V in Equation (5.5)], ranging from – 0.50 to 1.33. He provides two illustrations of the impact of such adjustments on utility, the first being negligible, the second substantial.

Schmidt et al. (1982) estimated the utility of training programmers as $1,303.69 [using Equation (5.2), with N = 100, T = 2 years, d_t = 0.65, SD_y = $10,413] (p.335). Boudreau assumes, for his illustration, that the net effects of positively and negatively correlated variable costs produce a V equal to – 0.05; the after-cost utility estimate based on SD_y of $9,892 [i.e., ($10,413) (1–0.05)] would be very similar to the original Schmidt et al. utility computed of $1,235,960, or 94.8% of the reported utility estimate.

However, for the Cascio and Sibley (1979) study, Boudreau found that Equation (5.2) would produce a large bias because high variable costs for the job of sales manager would be omitted from consideration. (Cascio and Sibley's purpose had been that of estimating the utility gain brought about by replacing an interview selection procedure with an assessment center selection procedure.) In their study, SD_y was estimated using the "dollar value of sales to the company" of sales managers who were probably paid, at least in part, on commission or incentive. Boudreau therefore assumes (although no data were provided in the original study) that V is equal to – 0.40 in this situation. The SD_{nb} value would then be $5,700 rather than the $9,500 reported in the original study; the adjusted utility estimate would be $87,782 rather than $153,835, or only 57% of the reported value.

Taxes, the second economic consideration, like variable costs are often an unavoidable obligation (except for the government and some other exceptions). Taxes assessed on profits produce a proportional reduction in both revenue and costs.

Boudreau (1983a) argues that accounting for taxes in personnel program utility is important so that investment in such a program can be compared to other investment options. He notes that a basic principle in capital budgeting analysis is that after-tax costs and benefits are the appropriate basis for decisionmaking. Also because organizations vary in their tax liability, inter-organizational utility comparisons require utility values be adjusted to account for different tax consequences. The marginal tax rate (the tax rate applicable to changes in reported profits generated by a decision) is the appropriate adjustment for dealing with effects of productivity "as sold" and on costs.

Boudreau notes that net benefits (nb) was earlier defined as $(1+V)$ (sv), and if marginal tax rate equals TAX, then after-tax benefits may be denoted $(1-TAX)$ (nb). Equation (5.5) can be rewritten:

$$\Delta U = (N)(T)(r_{x,sv})(\bar{Z}_x)(SD_{sv})(1+V)(1-TAX) - (C)(1-TAX) \quad . \qquad (5.6)$$

Taxes produce a proportional decrease in SD_{nb} and in C, usually reducing ΔU, because SD_{nb} is usually greater than C.

Boudreau suggests that the higher an organization's marginal tax rate, the lower utility will be, all else being equal. He assumes, for federal and state taxes, a range of marginal tax rates (TAX) from 0% to 55% which implies an adjustment [$1-TAX$, in Equation (5.6)] ranging from 1.0 to 0.45.

Boudreau turns again to the Schmidt et al. (1982) and Cascio and Sibley (1979) data to illustrate the combined effects of variables costs and taxes. Although the Schmidt et al. estimate of SD_y was derived for the federal government, Boudreau's analysis generalizes results to private-sector taxable organizations by including both federal and state tax rates. Assuming a marginal tax rate of 45%, and variable costs of 5% (as in the previous illustration above), the SD_y estimate would be further reduced from \$9,892 to \$5,415, but, at the same time, the estimated treatment cost of \$500 per person corresponds to a reduced after-tax cost of \$275 per person. Substituting these values into Equation (5.6), along with other parameters given in Schmidt et al., yields an estimate of \$676,450 rather than the reported \$1,303,690 for one application of the training program, or 52% of the original value.

For the Cascio and Sibley (1979) study, Boudreau assumes a 45% tax rate in addition to the 40% variable cost level assumed above. The yearly SD_y estimate, after variable costs and taxes, would be \$3,135 rather than the reported \$9,500. The differences

in cost between the assessment center and the interview would also be reduced from $11,328 to $6,230, after taxes. The resulting total payoff from assessment center selection procedures (holding the other parameters at the levels noted) would be $48,280--only 31% of the reported value of $153,855.

With regard to discounting, the third economic consideration, Boudreau (1983a) points out that where costs and benefits accrue over time, the value of future costs and benefits must be discounted to reflect the opportunity costs of returns foregone. Future monetary values cannot be equated with present monetary values, because benefits received in the present or costs delayed into the future would be invested to earn returns. Thus, a dollar received in 1988 at 6% annual returns would be worth $1.12 in 1990 and a future benefit worth $1.12 in 1990 has a 1988 value of $1.00 ($1.12/1.062). Boudreau derived the following utility formula to take into account discounting and the other two economic factors:

$$\Delta U = (N)\{\sum_{t=1}^{T}[1/(1+i)^t]\}(SD_{sv})(1+V) \times (1-TAX)(r_{x,sv})(\bar{Z}_x) - C\,(1-TAX) \quad (5.7)$$

where ΔU is the change in overall worth or utility after variable costs, taxes, and discounting; N is the number of employees selected; t is the time period in which an increase in productivity occurs; T is the total number of periods (e.g., years) that benefits continue to accrue to an organization; i is the discount rate; SD_{sv} is the standard deviation of the sales value of productivity among the applicant or employee population; V is the proportion of sales value represented by variable costs; TAX is the organization's applicable tax rate; $r_{x,sv}$ is the validity coefficient between predictor (x) and sales value utility; and C is the total selection cost for all applicants.

Boudreau further modifies the Schmidt et al. (1982) study estimates to take discounting into consideration. It was shown above that the combined effects of variable costs and taxes reduced the original SD_y estimate from $10,413 to $5,415. To illustrate the effects of discounting, Boudreau notes that the proposed duration effect used in the original study was 2 (the results in one-year payoff for two years). If the discount rate is assumed to be 10%, Boudreau computes the yearly payoff of $5,415, multiplied by 1.74, instead of 2; the after-tax utility estimate of the training is $584,937 rather then $1,303,690, or 45% of the original values.

Boudreau notes that in the Cascio and Sibley (1979) study, instability in expected performance was accounted for by the authors by assuming a correlation of 0.70 between any pair of periods. This adjustment had the effect of multiplying the yearly SD_y estimate by 4.36 (i.e., $\sqrt{19}$, rather than by the value of T (5 in this case). According to Boudreau this instability adjustment has the same effect as assuming a discount rate of 4.75%. Given the earlier assumptions, and using the same discount rate of 10%, the resulting total estimate after subtraction, for all three economic factors, for replacing an interview selection procedure with an assessment center procedure would be $41,154 rather than $153,835, or 27% of the original value.

Boudreau (1983a) concludes that economic considerations, individually and in combination, indicate substantial biases (especially upward biases) in published estimates when the estimates do not include the effects of variable costs, taxes and discounting. Equation (5.7) provides a more complete utility formula than those previously used and one that is likely to produce less biased (though often lower) utility estimates for personnel programs.

Boudreau further suggests that estimates of the tax rate and discount rates are already frequently used by organizations for various decisions and thus reasonably accurate estimates of these parameters may be available for research use. Estimates of variable service costs, however, are probably less readily available.

The value in using estimates that recognize economic considerations, Boudreau asserts, is that they may provide a more defensible and realistic utility definition. While Boudreau's definition can often lead to lower utility estimates, such estimates remain substantial and still provide compelling evidence of the value of personnel programs.

As we shall see in the next section, while failure to consider the combined effects of economic factors may overstate utility estimates considerably, the failure to consider the effect of employee flows may underestimate utility estimates even more.

Hunter et al. (1988), however, provide examples indicating that capital budgeting and financial accounting techniques advocated by Boudreau (1983a, 1983b) and Cronshaw and Alexander (1985) may be conceptually and logically inappropriate and conclude that there is no single correct definition of utility.

2. Effects of Employee Flows in Utility Analysis

Boudreau (1983b) extends previous utility models beyond reflecting the effects of personnel on one cohort of employees by incorporating the flow of employees into and out of the work force. Most early utility models assume that a selection program is applied to one group of applicants, and provide the utility of adding the one-treated cohort to the existing work force. In practice, of course, selection programs are administered over a period of years as employees flow into and out of the work force.

The traditional one-cohort model applies selection program effects to entering employee cohorts of size N and the program's benefits are reflected in that cohort group for T periods. But a program's effects on subsequent cohorts will occur in addition to its lasting effects on previously treated cohorts. Boudreau calls such effects, "additive cohort effects," and failure to consider them may substantially understate the program's utility.

Boudreau notes that employee flows generally affect utility through period-to-period changes in the number of treated employees in the work force. The number of treated employees in the work force--k periods in the future (N_k) may be expressed as follows:

$$N_k = \sum_{t=1}^{k} (N_{a_t} - N_{s_t}) \tag{5.8}$$

where N_{a_t} is the number of treated employees added to the workforce in period t, and N_{s_t} *is* the number of treated employees subtracted from the workforce in period t. The term N_k reflects both the number of treated employees in previous periods and their expected tenure.

The formula for the utility, ΔU_k, occurring in the *kth* future period that includes the economic consideration of Equation (5.7), may be stated:

$$\begin{aligned} \Delta U_k = & \left[\sum_{t=1}^{k} (N_{a_t} - N_{s_t} \right] \left\{ \left[1/(1+i)^k \right] (r_{x,sv}) \right. \\ & \left. \text{x } (\bar{Z}_x)\, (SD_{sv})\, (1 + V)\, (1 - TAX) \right\} \\ & - C_k\, (1 - TAX)\, [1/(1+i)^{(k-1)}] \end{aligned} \tag{5.9}$$

Boudreau notes that for the sake of simplicity, the utility parameters $r_{x,sv}$, V, SD_{sv}, and TAX are assumed to be constant over time. This assumption is not critical to this utility model, even though the factors may vary. The cost of treating the N_{a_k} employees

added in period k (C_k) is also allowed to vary over time. C_k is not treated simply as a constant multiplied by N_{a_k}. Also, the discount factor for costs $[1/(1+i)^{(k-1)}]$ reflects the exponent $k-1$, assuming that such costs are incurred one period prior to receiving benefits. Where costs are incurred in the same period in which benefits are received, k is the proper exponent.

Boudreau then states that to express the utility of a program's effects over F periods, the one-period utility estimates (ΔU_k) are summed. Thus the complete utility model reflecting employee flow through the work force for a program affecting productivity in F future periods may be written:

$$U = \sum_{k=1}^{F}\left[\sum_{t=1}^{k}(N_{a_t} - N_{s_t})\right]\left\{\left[1/(1+i)^k\right](r_{x,sv}) \times (\bar{Z}_x)(SD_{sv})(1+V)(1-TAX)\right\} - \sum_{k=1}^{F}\left\{C_k(1-TA\Xi)\left[1/(1+i)^{(k-1)}\right]\right\} \quad . \tag{5-10}$$

The duration parameter F in Equation (5-10) is not simply a function of employee tenure, but also depends on how long a program is applied. For example, Boudreau assumed that the PAT evaluated by Schmidt et al. (1979) is applied for 15 years or 5 years after the first cohort separates. If 618 programmers are added each year, then for the first 10 future periods N_k will increase by 618 in each period.

In Table 5.1, Boudreau illustrates the effect on selection utility using the Schmidt et al. data described early: $N = 618$; average tenure of selection (T) = 9.69 years; the validity of the PAT(r_{x_y}) = 0.76, and SD_y = \$10,413. Boudreau provides an explanation of the columns of the table and results:

> The first column of the table contains the future period (k) in which the utility occurs. The second column contains the number of treated employees (in this case selected using the PAT) in the work force in the kth future period (N_k). The third column contains the one-period gain in dollar-valued productivity (in the kth future period) from the N_k treated employees. The fourth column contains the discount factor applicable to G_k in the kth period [i.e., $DF(G_k) = 1(1+i)^k$]. The fifth column contains the discounted present value of G_k [i.e., the product of G_k and $DF(G_k)$]. The sixth column contains the after-tax cost of testing 1,236 applicants from which 618 new employees are selected. (Note that the device is applied

Table 5.1. Empirical Illustration of Employee Flow Effects on Utility

k	N_k	G_k [a]	$DR(G_k)$	$PV(G_k)$ [a]	$C_k(1\text{-}TAX)$	$DR(C_k)$	$PV[(C_k)(1\text{-}TAX)]$	ΔU_k [a]
1	618	$2.04	0.91	$1.86	$6,798	1.0	$6,798	1.85
2	1,236	4.09	0.83	3.39	6,798	0.91	6,186	3.38
3	1,854	6.13	0.75	4.60	6,798	0.83	5,642	4.59
4	2,472	8.18	0.68	5.56	6,798	0.75	5,099	5.55
5	3,090	10.22	0.62	6.34	6,798	0.68	4,623	6.33
6	3,708	12.27	0.56	6.87	6,798	0.62	4,215	6.87
7	4,326	14.31	0.51	7.30	6,798	0.56	3,807	7.30
8	4,944	16.35	0.47	7.68	6,798	0.51	3,467	7.68
9	5,562	18.40	0.42	7.73	6,798	0.47	3,195	7.73
10	6,180	20.44	0.39	7.97	6,798	0.42	2,855	7.97
11	6,180	20.44	0.35	7.15	6,798	0.39	2,651	7.15
12	6,180	20.44	0.32	6.54	6,798	0.35	2,379	6.54
13	6,180	20.44	0.29	5.93	6,798	0.32	2,175	5.93
14	6,180	20.44	0.26	5.31	6,798	0.29	1,971	5.31
15	6,180	20.44	0.24	4.91	6,798	0.26	1,767	4.91
16	5,562	18.40	0.22	4.05	0	0.24	0	4.05
17	4,944	16.35	0.20	3.27	0	0.22	0	3.27
18	4,326	14.31	0.18	2.58	0	0.20	0	2.58
19	3,708	12.27	0.16	1.96	0	0.18	0	1.96
20	3,090	10.22	0.15	1.53	0	0.16	0	1.53
21	2,472	8.18	0.14	1.14	0	0.15	0	1.14
22	1,854	6.13	0.12	0.74	0	0.14	0	0.74
23	1,236	4.09	0.11	0.45	0	0.12	0	0.45
24	618	2.04	0.10	0.20	0	0.11	0	0.20
25	0	0	0.09	0	0	0.10	0	0

Source: Boudreau (1983b), p. 400.

Note: Total utility is $105.01 million.

[a] Values expressed in millions.

only in periods 1 to 15.) The seventh column contains the discount factor applicable to the after-tax testing cost [i.e., $DF(C_k) = 1/(1+i)^{(k-i)}$], and the eighth column is the discounted present value of the after-tax testing cost [i.e., the product of (C_k), $(1 - TAX)$ and $DF(C_k)$]. The last column contains ΔU_k, computed as the difference between columns five and eight.

. . . The second column of Table 5.1 illustrates the effects of employee flows. For the first 10 future periods, N_k rises by 618 in each period. This is because the first selected cohort remains in the job for 10 years, and each newly hired cohort of 618 employees (assuming a constant number of selectees per period) is added to the first (i.e., $N_{s_t} = 0$). Beginning in future period 11, one treated cohort leaves in each period (i.e., $N_{s_t} = 618$). However, by continuing to apply the PAT to select 618 new replacements, the employee inflow is maintained (i.e., $N_{s_t} = 618$). Thus, in future periods 11 through 15, N_{a_t} and N_{s_t} offset each other and N_k remains unchanged. Beginning in future period 16, the PAT is no longer used (C_k and N_{a_t} become zero, assuming the organization returns to random selection). However, the treated portion of the work force does not immediately disappear. Earlier-selected cohorts continue to separate (i.e., $N_{s_t} = 618$), and N_k falls by 618 each period until the last treated cohort (selected in future period 15) separates in future period 25.

The total expected utility of the 15-year selection program is the sum of the last column, $105.1 million. This utility value is substantially higher than the Schmidt et al. (1979) utility estimate of $37.6 million, even reflecting variable costs, taxes, and discounting. (pp. 399-400)

Boudreau cautions that it may be tempting to conclude that the actual dollar payoff from valid selection programs is two or three times higher than one-cohort models predict. However, such a conclusion assumes greater precision in reported utility values than is presently justified, since all existing models contain parameters that must be estimated and, moreover, use simplifying assumptions that may be unrealistic, (e.g., assuming variable costs, SD_y, and the selection ratio are constant over time). While Equation (5.10) permits these parameters to vary, more data are needed on parameter accuracy and variability.

Boudreau concludes that employee flows through the work force not only shows that utility values are most likely substantially higher than estimated, but also highlights the importance of examining assumptions of stability in utility parameters over time. In particular, he suggests the development of a utility employee flows model that can more realistically consider the tradeoff between increased turnover and increased productivity. Accordingly, Boudreau's initial employee flows model provides a framework for future research aimed at linking utility concepts to a number of human resource areas such as planning, recruitment and employee turnover.

The integration of recruitment strategies into selection utility models is considered next (Boudreau & Rynes, 1985), followed by an external employee movement model (Boudreau & Berger, 1985) that permits even more realistic decisionmaking.

Boudreau and Rynes (1985) provided a more complete staffing utility model by integrating the effects on utility of recruitment activities that precede selection. Previous research assumed that the applicant pool represents a random sample from the applicant population. When this is so, the sample means and standard deviations are unbiased estimates of these values in the population. Applicant populations have been assumed to be constant; recruitment has been treated as a way to increase the size of the applicant pool.

But recruitment practices may alter the characteristics of the applicant pool. The authors note that Alexander, Barrett, and Doverspike (1983) argued that self-selection and initial organizational screening might cause "examinees" to be a non-random sample from the "population of interests." Boudreau and Ryne's conceptual framework addresses that observation by differentiating the parameters that characterize the applicant pool from the applicant population. Their focus is on the differences in the applicant pool generated by different recruitment strategies rather than on pretest score levels resulting from pre-screening.

Boudreau and Rynes (1985) introduced three changes in Equation (5.10) to reflect the effects of recruitment: the cost of recruitment is added to the cost parameter; utility is expressed in absolute, rather than incremental terms; and the equation permits possible parameter variations in dollar-value performance, the validity and in the selection ratio.

To demonstrate how differences in recruitment strategies affect the results of selection utility, the authors conducted a hypothetical utility analysis, using empirical parameter values from published studies, where possible. In the analysis, the authors assumed a job that has 10 vacancies per year and uses an on-site interview as the selection device. All utility parameters were assumed and inserted in the modified form of Equation (5.9). Two recruitment options were considered: using a private employment agency versus advertising plus resumé screening.

The results in the authors' illustration show that the selection-plus-recruitment utility produces different conclusions than would have been reached with conventional selection utility models. The incremental selection utility of the private agency option is $9,306 versus $99,091 for the advertising and screening option. This result, the authors

note, was obtained because of the wider variability and lower selection ratio for the applicant population in the latter recruitment option.

In contrast, the *combined* total recruitment and selection utility was $1,231,451 for the private agency option and $706,406 for the advertising and screening option. The utility values in this example reflect the advantage of the first option in producing a higher difference between the average level of service value and service costs per person selected. Improved recruitment raises the average value of the applicant population by enough to compensate for the reduction in the validity, variability, and selection ratio associated with that population.

Boudreau and Rynes conclude that decisions can be improved by accounting for recruitment utility effects, and emphasize the need for additional recruitment research using employee flows models that integrate recruitment effects.

Boudreau's (1983b) employee flows model was the first effort to improve the realism of the traditional one-cohort selection utility model. However, as noted by Boudreau and Berger (1985), this initial flow model assumes a specific pattern of acquisitions and separations in which the quantity and quality of selected employees are assumed to be equal. Thus the 1983 employee flows utility model represents a special case of a more general external employee movement model.

Boudreau and Berger (1985) proposed that utility models for employee selection decisions offer a useful framework for examining the utility implications of employee separations and for integrating selection and separation utilities. Both procedures affect the quantity, quality, and cost of acquisitions. The authors developed an "external employee movement" utility model that describes the consequences of employee movement out of and into the organization (separations and acquisitions).

Their model embraces three related types of movements: repeated acquisitions without separations; repeated unreplaced separations over time; and repeated separations over time replaced with new employees. The latter model, the authors state, is the most general case and provides an explicit link between separation utility and existing selection utility models.

The authors present the conceptual model graphically to show the close parallel between selection utility and separation (or retention) utility, and the important integrative relationship between the two concepts. Utilizing these concepts, they developed a complex algebraic derivation of a general utility model for external employee movement.

They show that their general employee movement utility model is not only consistent with the traditional selection utility model and the model modified by economic considerations [Equations (5.1) and (5.7)] and the original employee flows model [Equation (5.10)], but it considers and incorporates a larger set of important organizational activities. A central issue is whether the additional realism embraced by the new model can substantially affect decisionmaking.

Boudreau and Berger (1985) also reported the finding of a hypothetical situation that simulated various acquisition and retention strategies. They found that decisions on retention patterns could greatly affect utility. If, for example, no separation or acquisitions took place, the total work force value over a ten-year period would be $232.5 million. In this case, the organization would incur no employee movement costs, but would attain the movement benefits. They pointed out that retaining the most productive employees may produce a higher work force value, even without the use of a valid selection program. Work force value increases, as expected, when productive employees are retained, combined with high quality hires. If the validity of a selection program is 0.50, retaining the most productive employees can result in a work force utility of $337 million; if the validity is 0.00, retaining the most productive employees can result in a work force utility of $310 million. These figures, the authors note, do not deduct the additional costs incurred to improve retention (e.g., higher employee compensation).

The authors assert that the omission of retention utility may bias selection utility estimates, and retention effects need to be directly incorporated in determining selection utility. They noted that a number of studies have defined the cost of acquisitions and separations (e.g., Abelson & Raysinger, 1984; Cascio, 1982; Cawsey & Wedley, 1979; Dyle & Keaveny, 1983; Flamholtz, 1974; Gaudet, 1960; Savich & Ehrenreich, 1976; and Ward, 1982). The external employee movement model suggests that these analyses must also estimate the benefits of retention and acquisitions. Additionally, analyses are needed to estimate the costs of implementing programs that affect the quantity and quality of retention (e.g., retirement inducements).

Boudreau and Berger's external employee movement utility model is clearly the most sophisticated means of estimating utilities developed to date. It provides a vehicle for even further integration and expansion to increase realism and accuracy--a model that integrates external and internal employee movement into a general utility analysis framework.

CHAPTER 6. CURRENT ISSUES IN UTILITY ANALYSIS

This chapter addresses several current issues in utility analysis:

- Limitations of utility analysis attributable to assumptions made in its applications;
- Controversies in estimating utility parameters;
- Linking human resource models to economic theory;
- Using classification decisions to improve human resources utilization.

A. LIMITATIONS OF UTILITY ANALYSIS: MAKING ASSUMPTIONS AND ESTIMATES

A decision problem arises whenever an individual is confronted with alternative courses of action. In the context of personnel interventions, focus is on institutional decisions where the most generally useful strategy is one that maximizes average gain or minimizes average loss. Cronbach and Gleser (1965) write:

> Decision theory is distinguished from simpler models by the fact that it is built of concepts that are often neglected: the set of alternative treatments, the costs of experimentation, the possible outcomes and the payoffs associated with them, etc. Yet when one seeks to make use of decision theory, he almost invariably sets a number of these key concepts aside, so as to make the model tractable. It makes sense in certain problems of sequential test design, for example, to assume the cost of collecting data to be negligible; to bring in such costs explicitly makes the problem much harder to think about. On the other hand, the *concept* of cost is indispensable, since, if cost were truly negligible, it would never be advantageous to terminate data collection.
>
> There are two levels at which decision theory is to be comprehended. It is, on the one hand, a set of notions, all of which are to be kept in mind in formulating questions. Second, it is a formal machinery for determining optimum strategies. In working out formal solutions it is invariably necessary to neglect certain of the key concepts, to introduce strong assumptions, or to ask for detailed information that cannot practically be supplied. Even when one is using decision theory in only a "notional" way, it is often necessary to simplify the model to keep the discussion within bounds. (p. 151)

It is clear, then, that an advantageously distinguishing feature of utility analysis is the degree of realism it contains compared to simpler models. All utility analyses make simplifying assumptions directly or indirectly and also make parameter estimates that cannot be measured precisely. Accordingly, the validity (accuracy) of utility analysis findings always depends on the considerations embedded in the analysis, and on the assumptions and estimates made.

Although the basic utility model encompasses many of the most significant parameters likely to change decisions, it is limited by the information considered and the assumptions and estimates made. Boudreau (1983a) removed some of these limitations by extending the basic model to include economic considerations, employee movement flows and risk analysis. Schmidt et al. (1982) extended the generalizability of the basic model to all interventions designed to enhance performance.

It is instructive to examine the assumptions contained in the basic utility equation since all recent variations and extensions are based on Brogden's original formulation. Cronbach and Gleser list seven basic assumptions:

> 1. Decisions are made regarding an indefinitely large population of persons. This "*a priori* population" consists of all applicants after screening by any procedure which is presently in use and will continue to be used.
>
> 2. Regarding any person i, there are two possible alternative decisions: accept (t_A) and reject (t_B).
>
> 3. Each person has a test score y_i, which has zero mean and unit standard deviation.
>
> 4. For every person there is payoff e_{it} which results when the person is accepted. This payoff has a linear regression on test score. The test will be scored so that r_{ye} is positive.
>
> 5. When a person is rejected, the payoff e_{it} results. This payoff is unrelated to test score, and may be set equal to zero.
>
> 6. The average cost of testing a person on test y is C_y, where $C_y > 0$.
>
> 7. The strategy will be to accept high scoring men in preference to others. A cutoff y' will be located on the y continuum so that any desired proportion $\phi(y')$ of the group falls above y'. Above that point probability of acceptance is 1.00; below it, 0.00. (p. 1-1, Appendix)

The concept of cost was assumed to be a central consideration in decision theoretic models from their earliest inception; costs were to be subtracted from savings attributable to

selection procedures to arrive at net benefits (or productivity gain). Brogden (1949) and Cronbach and Gleser (1965) addressed themselves principally to testing costs and suggested heuristics for limiting such costs to optimize gains. Hunter and Schmidt (1982) appropriately assumed that testing costs (e.g., for paper-and-pencil tests) were relatively insignificant compared to productivity gains in some contexts.

In recent years, investigators striving for greater precision in utility estimates gave careful consideration to cost factors. In costing assessment centers, Burke and Frederick (1986) include costs of setting up the assessment center; consulting fees to maintain the assessment center; assessor training costs; salaried time away from sales territory for both assessors and assessees; materials; and travel hotel and meal costs for assessors and assessees. Other costing examples are variable costs, taxes, and discounting (Boudreau, 1983a); recruiting costs (Boudreau & Rynes, 1985; Fernandez & Garfinkle, 1985); employee movement acquisition and separation costs (Boudreau & Berger, 1985); and costs of developing and conducting a training program (Mathieu & Leonard, 1987).

Cronbach and Gleser's (1965) first assumption stipulates that selection procedures should be evaluated in the situation in which they are to be applied: selection utility estimates should be based on the appropriate applicant population. In practice, however, the incumbent population is used to obtain SD_y estimates from supervisors since supervisors are most familiar with the performance not of applicants but of their employees. Estimates based on the incumbent population may be inaccurate because the values estimated for employees tend to be restricted compared to those of applicants and thus are believed to be conservative (Schmidt et al., 1979); the values based on an incumbent group may not be representative of future groups because labor market conditions may change or recruiting strategies are not uniform (Boudreau & Rynes, 1985); and supervisors may fail to take into account the effects of varying tenure and job duties of incumbents in making SD_y estimates--use of incumbents as surrogates for applicants where these problems would not arise (Bobko et al., 1983; Boudreau & Rich, 1986).

Additionally, the payoff functions assumed in analyses differ from one investigator to another and hence are not defined in similar economic units (Boudreau, 1984). For example, as noted earlier, utility studies defined SD_y as the "value of sales" (Cascio & Silbey, 1979); the "value of products and services" (Schmidt et al., 1979); or "what the employer charges the customer" (Hunter & Schmidt, 1982). Boudreau (1983a) defined the payoff function as net benefits--the difference between sales value and service costs.

As noted in the previous chapter, the measurement of the payoff scale has received the most attention. Concern with the variability and accuracy of SD_y persists; although several techniques have been suggested to improve the reliability and credibility of SD_y, controversy still surrounds its measurement, and no standard estimating method has been universally accepted.

Boudreau and Berger (1985) distinguish between parameter estimates made in previous studies and new types of parameter estimates required for their sophisticated utility model of external employee movement. The payoff scale is one of several parameters in utility equations requiring estimates.

Parameters that were used in previous utility research relating to incremental acquisition utility include the number of acquisitions in future periods, the validity coefficient, the average standardized selectee predictor score, the dollar-valued standard deviation of service value among applicants, costs of activities associated with the proposed selection method, and economic and financial parameters (variable costs, taxes, and discounting).

Incremental utility parameters related to nonrandom retentions not previously used include the number of separations, the transactions cost of the separations, and the measure of the service value difference between incumbents and retainees (similar to performance difference between leavers and stayers).

Boudreau and Berger also describe service value and service cost estimates that are new to the employee movement model that reflect average incumbent service value and service cost over time.

A final parameter estimated in their analysis is duration effects, e.g., the model is "selecting" incumbent employees, and the best estimate of the duration of the effects would be the expected tenure of retained employees.

Boudreau and Berger (1985) evaluated the appropriateness of six "traditional" assumptions pertaining to the treatment applied to the one-cohort acquisition model, Equation (5.10). They noted if they applied all of the same assumptions, their expanded employee movement model would produce precisely the one-cohort model and consequently would fail to encompass and integrate a much larger set of important organizational activities.

The authors noted, however, some of the assumptions in the one-cohort model limited selection utility to a relatively narrow subset of possible decision situations, even though these were not unrealistic. Other assumptions were more difficult to accept, and may also have been irrelevant to decisionmaking. Consequently, the authors decided to remove several assumptions and retain several others in their new external employee movement model.

The examples given above demonstrate how recent expanded utility formulations strive for more precision by incorporating additional concepts, and by using more realistic assumptions and parameter estimates. Such enhanced utility formulations should contribute to greater credibility and acceptance in organizational decisionmaking.

B. LINKING HUMAN RESOURCES MODELS TO ECONOMIC THEORY

The determination and interpretation of productivity gains resulting from personnel interventions are consistent with the economic way of thinking and of making decisions. There are, however, some caveats. The commonly accepted basic economic definitions we use here are taken from Heyne's (1988) microeconomics text.

Opportunity cost, Heyne states, is a concept that ties together the law of demand and the principles governing supply. Economists think of cost as the value of sacrificed opportunities. The real cost of any action is the value of the best alternative opportunity that must be sacrificed in order to take action. A clear example of the opportunity-cost concept is the value placed on land. The price an individual pays for land depends on the value of that land in alternative use.

The concept of opportunity costs explains how labor enters into production costs. Employees must receive from their employers compensation that convinces them to turn down all other job offers. A skilled worker is paid more than an unskilled worker because and only insofar as the skilled worker's skills make him more valuable somewhere else.

According to Heyne, the economic way of thinking recognizes no objective costs. For example, the cost of a volunteer military force must consider an enlistee's opportunity cost as foregone alternative employment opportunities as well as other values (e.g., life-style preferences, attitude toward war). When the military bids for recruits, it raises its offer to the point that it can attract the desired number and quality of enlistments. In a sense, the military is actually trying to minimize costs by attracting those with the lowest opportunity costs of service and still meet its need. The cost of a conscript force will

almost certainly be greater. If more draftees come from the upper quality portion of the supply curve, the higher the cost will be to conscriptees. The government, in a draft, transfers that cost from the shoulders of the taxpayers to the shoulders of the draftees (Heyne, 1988).

The costs that influence supply are always marginal costs, or expected additions to the potential suppliers' costs. Economic theory is based on marginal analysis because it assumes that decisions are always reached by weighing additional costs against additional benefits. Nothing matters in decisionmaking except marginal costs and marginal benefits. Opportunity costs are always expected marginal costs. The term marginal costs does no more than bring into strong relief an aspect of opportunity-cost thinking.

Irreversible costs made before a decision are called by economists "sunk" cost. Sunk costs are irrelevant to decisionmaking because they represent no opportunity for choice. They may be cause for regret, but are no longer relevant to the economics of present decisions.

Sometimes the marginal concept is confused with the notion of average. The incremental expenditure in producing the last item (or batch of items) of a product, is the marginal cost of the item. The marginal cost can be more or less than the average cost of producing all of the items. However, businessmen generally do not commit themselves to a course of action unless they anticipate being able to cover their total costs. Managers may set up problems in terms of anticipated production costs per unit against anticipated selling price per unit, but anticipated costs of any decision are really marginal costs. It is expected marginal costs that guide decisions: economic decisions are always made in the present with an eye to the future. Thus to maximize profits or minimize loss (which is really the same thing), the anticipated marginal cost of a product must be equal to the anticipated price at which the firm sells in a competitive market.

In microeconomic theory the relevant cost to an organization or to an individual is the value of whatever is given up through the decision--the marginal cost (or opportunity cost). An increase in the demand for any good will bid up the cost (or price) of acquiring the good, to the extent that it does not cause a larger quantity to be supplied. Conversely, an increased demand for any good will not raise its price to the extent that suppliers respond by making larger quantities available. A good example of how labor enters into production is found in the job of artificial intelligence (AI) researcher. Capable AI researchers were in great demand and short supply during the 1970s and thus were paid

relatively well compared to other researchers. In 1988 demand was still great, but the supply larger; AI researchers were still being paid well, but relatively not as well as in earlier years.

The response of the supplier will depend on the marginal cost of transferring resources out of their current uses into the production of the good for which the demand has increased (e.g., the AI researcher estimates the marginal benefit of staying at his university job versus accepting a job in industry).

In determining economic marginal values or utilities of a production function, a broad range of relevant organizational costs enter into consideration (e.g., labor, materials, equipment, financing, etc.). In selection utility, by way of contrast, productivity gains are estimated only for effects attributable to the selection program (i.e., increased productivity of higher quality workers), holding all other factors constant. In this sense, selection (quality of labor) may be considered as a single component of the marginal utility of a production function.

Production theory implies that a factor's marginal product is dependent on the relative amount of other factors with which it is combined. The same holds for measuring individual productivity. Although job performance measures link individuals with their jobs, actual productivity will be influenced by the number and quality of co-workers, by the quality of equipment available to do the work, organization-wide policies and practices, etc. Additionally, individual measures of performance will not provide the data needed to determine the best combination of production factors.

Although present utility models incorporate a number of organizational activities (e.g., acquisitions and separations), they still do not include a number of important interacting organizational phenomena (e.g., planning, compensation, promotion, etc.), nor do present models consider external labor market considerations impacting on the organization. The interactions among these internal and external factors affect the accuracy of estimates and the generalizability of potential productivity gains to an organization's future situation.

With regard to internal factors, utility estimates pertain to potential productivity gains of *future* productivity increments attributable to higher quality employees, in the context of the *present* organization. We cannot estimate the combined effects of higher quality employees with other functional areas or the interaction of selection with any one area. Successful human resource programs are intended to bring about compositional

(quality mix) changes in the work force or changes in the characteristics of employees (e.g., change brought about by training, human relations skills programs, motivational programs, etc.). Hopefully, composition and characteristic-changing program effects would impact positively on other functional areas, including changes in the nature of some jobs or the critical task dimensions within a job. Such changes may impact on dollar-valued performance and other utility parameters that may not remain stable.

Because personnel programs (and other technological innovations) affect not only blue-collar and entry-level jobs, but increasingly affect white-collar and managerial jobs, the organizational situation can appropriately be characterized as ever-changing. Like economic decisions, personnel intervention decisions are made with an eye to the future. Better test-selected employees, assumed to be entering the future organization, will increase productivity; all other parameter estimates in the future organization, however, are assumed to be the same as they were in the present organization. Thus, since our current models neither incorporate all important organizational phenomena nor estimate interaction effects, present utility projections of a single production function (e.g., selection), may not accurately reflect utilities made for that same function at a later date.

Additionally, the effects of a personnel intervention are based on aggregating performance measures across all individuals performing the same organizational job. However, much of present organizational behavior concepts and practices focus on the performance of groups or teams. Most managers (and coaches) will assert that team performance is not the sum of the performance of individuals. In some types of work situations, team performance demands may sharply influence the quality/quantity of outputs produced by others in the group.

The relationship between group output and the mix of members' abilities in the workplace is largely unexplored. Knowing more about this relationship also would help to determine how to spread high quality individuals across organizational elements. Most managers will agree that individuals who contribute to overall effectiveness of their teams are especially valuable to organizations. Typically, the synergistic impact of such individuals increases the value of all other members, even those having average skills (Eaton et al., 1985).

It is common practice in the Army, for example, to use the concept of synergism; commanders assert that the best combat teams are those led by "leaders" of average soldiers. The leader ensures that each team member knows his job and is motivated to do

it. If this is true, it is clearly more cost-effective to assign a "leader" to each team and then comprise the remainder of the team with "average" soldiers rather than to assign only leaders to comprise an entire team. Aggregating productivity gains based on individual performance may be more appropriate in those situations where output is largely dependent on individuals working as separate entities (e.g., salespersons). In brief, utility analysis relies on individual measures, but organizational outcomes often depend more on the work-group level.

External to the organization, labor market conditions significantly impact on the pattern of acquisition and retention of employees. They directly affect the supply of high quality employees, compensation and other organizational inducements needed to retain them. Utility estimates pertaining to the characteristics of the applicant group and costs of employees change with labor market changes. Moreover, opportunity costs paid to valued employees largely reflect only that portion of a worker's economic value that is easily transferable from one employer to another (Becker, 1964). Employers who require only "firm-specific" employee skills are relatively free from having to bid against other firms for their own employees; otherwise they would have to take action to counter better offers tendered their employees by competing firms. However, by definition applicants for jobs requiring firm specific employer skills do not possess those skills; they must obtain such skills through on-the-job training or in-house promotions. The use of effective selection measures is particularly important because the firm must invest significant costs in training employees. Finally, it is unlikely that any production function will continue to increase linearly with ever increasing increments of high quality employees over an extended period of time; the law of diminishing returns eventually applies. If there are too many high quality workers, they inevitably will be assigned to tasks of less value and affect utility estimates accordingly.

Greer and Cascio (1987) assert that costing human resources has run aground due to the difficulties associated with operationally defining a relatively soft concept, the human worker. Admittedly, they state, there are difficulties in finding an acceptable method for valuing human assets as a balance-sheet item. However, Cascio (1987a) maintains that, contrary to common belief, all aspects of human resources management can be measured and quantified in the same manner as any operational function (Driessnack, 1979).

While utility analysis is the widely preferred strategy for estimating anticipated institutional gains from various courses of action, Cascio (1987a) outlines earlier human resource "asset" accounting procedures and the current "response" model. Cascio's

discussion of human resources accounting covers models that generated some excitement in the accounting field 15-20 years ago. Paperman (1977), however, concluded that the conservative principles of the accounting profession made it unlikely that human resources accounting would gain acceptance among accountants because of its reliance on subjective measures.

The *historical cost* approach to employee value (i.e., expenses actually incurred) is an asset model of accounting; it measures the organization's investment in employees. It is viewed as most appropriate for the purpose of external reporting (Tsay, 1977). The asset approach is relatively objective and it facilitates comparing levels of human resource investment on a constant basis with accounting treatment of other assets. It has a number of shortcomings, including the exclusion of any measure of value of employees to the organization.

Alternative asset measures noted by Cascio include the cost of replacing an employee (Flamholtz, 1971), economic valuation of average employees (Lev & Schwartz, 1971), and opportunity cost to the organization (Hekimian & Jones, 1967). The major limitation of human resource asset accounting models is that they focus exclusively on investments in people (inputs); they ignore effectiveness considerations as the output people produce.

Cascio notes an advancement in human resources accounting suggested by Pyle (1970) that properly compares input and output measures. However, it still fails to distinguish between individual and group effects that produce variability in output.

A conceptual approach in human resource accounting endorsed by Cascio (1987a) is termed an "expense" model (Mirvis & Macy, 1976). "Asset models assess the value of employees treating them as capitalized resources (i.e., the economic concept of human capital). In contrast, an organization uses human resource *expense models* to measure the economic effects of employees' behavior" (p. 6).

In the expense model approach, dollar estimates are attached to employee behavioral outcomes (e.g., turnover, absenteeism) typically found among workers in an organization. For example, in costing labor turnover, dollar figures are attached to separation, replacement, and training costs. Now this method does not measure the value of the individual but does take into account the economic consequences of the individual's behavior. Cascio asserts that expense measures--dollars--are taken quite seriously by most decisionmakers.

To repeat, the general idea of costing human behavior is not a new one, as evidenced by Brogden and Taylor's (1950) article calling for the development of on-the-job performance criteria in dollar terms.

Cascio (1987a) examined some important areas in which costs were attached to personnel activities, including turnover, absenteeism, and smoking. He computed the cost of smoking in the workplace at about $2,600 per person per year. Additionally, he computed dollar estimates for an intervention designed to improve job attitudes and also demonstrated costing considerations in labor contracts.

We close this section with an example of a promising economic model in the military context. Black (1987) pointed out that from an economic viewpoint, setting entry standards for enlistees as a means of improving productivity is analogous to maximizing the value of productive output by altering the quality of labor inputs, subject to a cost constraint. Individuals work together and with equipment to produce goods and services that contribute to national defense. The strategy considered is to change entry standards to maximize the value of enlistees' contributions to national defense. Increasing quality of enlistees continues until the marginal cost associated with enlisting a higher quality work force equals the marginal increase in the value of output, which in turn depends upon the *higher quality of that output, and possibly, higher volume of output. The expected change* in output is attributable to the performance-ability relationship.

The value of output depends on performance treated as a "physical" concept, (e.g., number of machines repaired properly per period), the value of the physical output, and values attached to output that weigh the relative importance of jobs (its shadow price) according to their contribution to defense, independent of individual performance. Other salient factors in such a modeling approach include expected career lengths (attrition, reenlistment rates), military discount rate, and costs of recruiting and training alternate ability groups.

Black noted that military changes in entry standards have multiple effects. They shift the ability mix of the force which affects job proficiency, and they alter patterns of attrition and reenlistment which affect duration of individuals' contributions, all of which need to be considered. He noted (and we agree) that if personnel could be allocated by means of an efficient classification and assignment procedure, productivity gains could dwarf the effects of changing entry standards on individual standards per se.

C. CLASSIFICATION DECISIONS AND HUMAN RESOURCE UTILIZATION

In this section we contrast major characteristics of selection and classification decisions and differential validity as a bridge to classification efficiency considerations that are addressed in detail in a subsequent report on measuring and improving classification benefits (Johnson and Zeidner, 1989). We also provide an overview of military selection and classification procedures as an introduction to existing manpower policies and evaluation techniques that are addressed in detail in a final report (Zeidner and Johnson, 1989).

1. Comparison of Selection and Classification Decisions

Selection tests are used to accept or reject an applicant for a job. Once an organization accepts an individual for employment, classification tests may be used to assign an individual to a specific training program or job from among a number of available opportunities. The purpose of classification is to match individuals and jobs in a manner that maximizes aggregate performance.

Classification decisions are a major concern in the military services and of increasing interest in industry. Classification also is used in counseling to provide guidance to students in the choice of a field of study or of an occupation, and in clinical diagnosis to aid in the choice of a course of treatment.

Placement tests may be used as an additional device in deciding on appropriate job levels in assigning individuals of varying qualifications, (e.g., placing more qualified individuals to higher than entry-level jobs). A job knowledge test may be used in placement within a skilled trade or an achievement test may be used in placement for studying a foreign language.

Traditionally in selection and placement decisions only a single job is involved, and can be accomplished with one or more predictors. The outcome is determined by an individual's position along a single predicted performance continuum. Classification, however, requires multiple predictors, jointly measuring more than one dimension. Validity is determined individually against each job's performance criterion, the set of job criteria should also be multidimensional. Thus a classification battery requires a separate least squares estimate (LSE) for each criterion. The particular combination of predictors employed out of the total battery, and the specific weight given each predictor, varies with each job criterion (Brogden, 1955; Horst, 1956a). In practice, a smaller number of tests

than are in the total battery are often used rather than the LSE, the complete regression equation for all predictors. In the Army, for example, a different unit-weighted, three-test combination or "aptitude area" composite currently is used in assigning individuals to jobs in each of nine job families.

It is often assumed that the utility of the classification process is a direct function of differential validity. More precisely, differential validity is the level of prediction using LSEs of differences among criterion scores. We will also use the terms in reference to the variation of a validity vector with a job having high differential validity, or being more valid for its own job family than any other job family. Unfortunately, a simulation study is required to translate the effect of differential validity into mean predicted performance (MPP) that in turn can readily be translated into utility. The utility of a classification battery can be characterized as being directly proportional to the average predicted performance of incumbents in a number of different jobs.

When the test content of the selection/classification battery has been fully determined and only the selection of the test composites and weights for use in the selection and/or classification of applicants for each job remains to be determined, the least squares regression weights applied to all tests forming each test composite, the LSEs, provide maximum utility when used in both selection and classification. Such composites will not only provide the means of maximizing the average validities across jobs but will also maximize potential allocation efficiency (PAE). The validities of these composites are, of course, the multiple correlation coefficients between the composites and each job criterion measure. No set of composites selected to lower intercorrelations among composites nor to increase the variations of composite validities across jobs (as one might mistakenly attempt to do in order to increase PAE), can increase the utility function value, as compared to the full regression equations based on the total battery. If composites have used a reduced number of tests or otherwise are not LSEs, the best composites for selection are not necessarily the best for classification.

A simple characterization of the concept involved in selecting tests or composites for classification purposes may be illustrated by considering two multidimensional jobs. In a two-job classification problem, the ideal composite to select from a yet undetermined battery of tests would be one that has a high correlation with the first job, and a zero, or still better, a negative correlation with the second job. Attaining differential validity requires identification of a predictor composite that is a good predictor for the first job and a poor predictor for the second job. A different predictor composite is then selected from that

yet undetermined battery of tests that is a good predictor of the second job and a poor predictor of the first job. Additionally, the lower the correlation between the two predictor composites selected for the classification battery, the better the differential validity.

A general mental ability composite, as is used to determine eligibility to enter military service, would be relatively ineffective in obtaining differential validity, since such a composite predicts performance across all jobs comparably well. Brogden (1946b, 1951, 1954) and Horst (1954, 1956b) independently developed the principal psychometric theories and methodologies used for selecting tests to improve the potential allocation efficiency (PAE) of a classification battery.

Brogden (1951) also presented a multidimensional one-stage selection/classification process that shows how to make much better utilization of human resources than is possible with either a single test or a single composite score from a regression equation. This is so because in assigning individuals to jobs, smaller selection ratios are used and consequently more qualified individuals are allocated to each job when multidimensional selection is used. For example, if out of 100 applicants, 10 were needed to fill each of two different jobs, the selection ratio is 0.10 for each job, when separate predictors are used for each. If a single predictor were used to select applicants for both jobs, the selection ratio would be 0.20, since at best the top 20 applicants need to be taken. Brogden developed a table that shows the increased efficiency resulting from the replacement of a single predictor with several LSEs, one for each job, when his assumptions are met. Even in the most extreme case, when all applicants are assigned to one of two jobs (SR = 0.50), assignment is random, and the correlation between the predictors is high (0.80), mean job performance still exceeds the chance value by 0.17 standard score units. As selection ratios become more favorable, mean performance gains resulting from classification become significantly higher: the mean standard criterion score increases from 0.17 to 0.96 when the selection ratio is 0.05.

Maier and Fuchs (1972) empirically demonstrated the pronounced advantage in using the Army's aptitude area composite system in assigning enlistees to specific jobs rather than using a single global general mental ability composite, the Armed Forces Qualification Test (AFQT). Anastasi (1988) notes that AFQT scores showed 56% of a sample of 7,500 applicants reached or exceeded the 50th percentile on AFQT, while 80% reached or exceeded the average standard score on their best aptitude area score. Thus a very large majority of enlistees assigned to jobs on the basis of aptitude areas would achieve predicted performance scores as high or higher than the average score of the entire

sample. This apparent impossibility in which nearly everyone could be above average, is attained by capitalizing on intra-individual differences for nearly everyone excels in some aptitude.

As noted earlier, the evaluation of a classification battery in meeting an organization's manpower policy objectives must always be considered together with the assignment process (i.e., the person-job matching procedure). A simulation study, employing an optimal allocation procedure, rather than the analytic method possible with respect to selection, is required to translate the effects of various strategies for improving classification (e.g., differential validity) into an improvement of MPP.

In 1949 the tests of the Army Classification Battery (ACB) were organized into aptitude areas, or combinations of tests for assigning individuals to various Military Occupational Specialties (MOS). The resulting classification system was a major innovation in military personnel utilization. When compared with the single measure for the Army General Classification Test of World War II, tests developed with differential classification in mind were shown to meet more total personnel requirements with better overall validity. In the old system, using a single measure of general mental ability, individuals with high scores would be assigned to jobs demanding complex cognitive skills, while individuals with low scores would be assigned to less complex jobs. In the new system using aptitude area scores, classification would be based on demonstration of specific cognitive ability composites necessary for a particular job while at the same time utilizing total human resources more efficiently. Thus aptitude areas allowed the use of scores that indicated differences in the levels of abilities and differences *among* abilities *within* each individual (inter- and intra-individual differences).

The value of using several aptitude areas, rather than one composite, depends, as noted, upon the presence of potential allocation efficiency (PAE) in the battery from which the tests comprising the aptitude areas were drawn. There was considerable PAE in the various versions of ACB during the first fifteen years of its use. Unpublished Army simulation studies showed a generally declining trend in the amount of PAE present with each change of ACB content during the period that the ACB was being transitioned into the Armed Services Vocational Aptitude Battery (ASVAB).

A serious shortcoming of the current ASVAB aptitude area composites for Army use is their inability to differentiate among job families. The same aptitude area used to select individuals specific to an MOS within a job family does nearly as well for MOS in

other job families. Each aptitude area is about as valid for other job families as it is for its own. More specifically, of the nine aptitude areas, only two are more valid for their relevant families than across families. Thus, while the operational composites are highly valid, the battery's composites appear to lack differential validity and one would expect to find a reduced amount of PAE in the ASVAB. (McLaughlin, Rossmeissl, Wise, & Brant, 1984; Zeidner, 1987).

Hunter, Crosson, and Friedman (1985) drew the same conclusion after analyzing occupational composite validities for job families in each service. At the time of the study there were nine families in the Army, four in the Air Force, five in the Navy and six in the Marine Corps. If different aptitude composites in actuality predict different job families, the validity of each occupational composite should have been highest for its own associated job family and lower for the other job families. Such a result would have been indicative of differential validity. The results, however, indicated that each occupational composite was almost as valid for other job families as it was for its own.

The conclusion we reach considering both studies, then, is that the ASVAB operational composites provide high validity but have little differential power as predictors of which assignment an individual should receive to maximize average performance for all jobs. With little PAE implied by the lack of differential validity (an approximate measure of classification efficiency), the benefits obtainable from using more than one occupational composite appear questionable.

It may be inferred that during the last two decades both test development and the selection of tests for inclusion in operational batteries have been directed toward the objective of maximizing the average validity of aptitude composites while ignoring the possibility that PAE might be lowered in the process. Nevertheless, it might be possible to exploit PAE in future operational batteries designed expressly for that purpose and also retain the conventional ability domains present in the ASVAB.

A more comprehensive study of the effect of test-criterion combinations on classification efficiency would include other predictor constructs. The ASVAB has been validated against carefully developed criterion measures of training, hands-on and job-knowledge performance measures, and performance ratings. The Army Research Institute developed such predictor-criterion measures and is currently carrying out a comprehensive validation study, called Project A (Eaton, Hanser, & Shields, 1986).

Campbell (1986) noted that Project A is guided by a view of job performance as being really multidimensional. He stated that "There is not one outcome, one factor, or one anything that can be pointed to and labeled as job performance. It is manifested by a wide variety of behaviors, or things people do, that are judged to be important for accomplishing the goals of the organization." (p. 7). In other words, the concept of total performance is more than technical proficiency; it includes contributing to teamwork, continuing self-development, supporting the norms and customs of the organization and persevering in the face of adversity.

The data provided in Project A indicate that differential prediction occurs across the major components of job performance. Differential predictors cannot be indicated in just the predictor or just the criterion space; differential prediction reveals itself only in the joint predictor-criterion space. Thus, there is differential prediction across the major components of the predictor universe. Further, the series of scaling studies conducted by Sadacca, deVera, DiFazio, and White (1986) show that judges can reliably indicate the relative importance of each criterion component within an MOS and that the patterns of weights differ by MOS. To the extent such differential utilities can be measured, they add to the PAE. It is obvious that the PAE cannot be other than zero if the criterion space is unidimensional.

If tests are selected with PAE in mind, and (given the multidimensional predictor and criterion space and hopefully a multidimensional joint predictor-criterion space along with differing utility of jobs by performance outcomes as reported in Project A), we may be able to resolve a major research issue. Campbell (1986) stated that ". . . one major research question we hope to answer is whether it is <u>ever</u> possible to estimate the parameters necessary for building a true classification algorithm. If it can't be done with a sample of 20 jobs and 500 cases per job then perhaps the textbook discussions of the classification problem are a bit academic." (p. 12). Even small increments in PAE (e.g., Harris, 1967) bring about worthwhile improvements in the military person-job match system.

Brogden (1959), as noted previously, showed that with other factors constant, and his assumptions met, the utility gain from classification varies with the intercorrelations among estimates of job performance by the function $\sqrt{1-r}$. Even when the correlation among the estimates is high, considerable utility remains, e.g., when $r = .80$, classification gains are 45% as great as with intercorrelations of zero. Although it has not yet been proven mathematically for the general case by Brogden (1959), it has been shown that the

higher the differential validity of each predictor composite, the higher the overall mean predicted job performance will be when an optimal personnel assignment model is used. We will show in a subsequent report that, under Brogden's assumptions, differential validity is important to MPP for a given number of jobs (Johnson and Zeidner, 1989).

There is a choice in the selection of tests for inclusion in the ASVAB: either the focus can be on validity generalization (a single general cognitive ability composite) or PAE can be improved using a technique such as Horst's stepwise selection procedure against multiple criteria (here meaning across different jobs); for a moderately large body of tests (e.g., ten or more) neither has to be improved at the expense of the other in the selection of tests. The addition of tests with high PAE does not need to detract from the validity generalization of a number of tests selected to maximize validity generalization, nor does the addition of tests with high validity generalization capability need to detract from PAE provided by tests specifically selected to maximize PAE.

Thus, it appears that the implementation of a selection/classification strategy that calls for selecting some tests to maximize the magnitude of validity coefficients and other tests to maximize PAE can achieve most of the PAE possible while losing little, if any, capability for validity generalization. Once a battery is selected, the same weights are best for achieving either the maximum average validity in accordance with validity generalization or the maximum PAE. Of course, the maximum PAE would not be achieved unless a maximum allocation procedure is used.

The possibility of fully benefitting from a deliberate consideration of PAE, with some decrease in average validity as a consequence, depends upon the following four conditions: (1) most important, whether the battery and composites are fixed (already determined); (2) whether the selection/classification process is accomplished in one or two stages (simultaneously or sequentially); (3) which optimal selection/classification procedure is being utilized to implement assignment to jobs (an LP type program); and (4) whether job families are appropriately structured (smallest LSEs in a job family and larger LSEs across other job families). The subsequent report (Johnson and Zeidner, 1989) details several ways classification efficiency can be improved.

In summary, it is apparently believed by many that the overall utility of a classification system is always improved by maximizing the average validity across all jobs even where doing so reduces differential predictability. It is also believed that utility is usually reduced by deliberately increasing PAE rather than focusing only on average

validity. Such beliefs are true only when the tests comprising the selection/classification battery have already been designated.

2. Challenges to the Differential Assignment Utility Concepts

In this section we discuss three major challenges to the concept of using a set of differentially valid test composites coupled with an optimal assignment procedure for matching military personnel to jobs. The first of these challenges concerns the use of cost-effectiveness measures in the military context; the second concerns the transformation of performance measures into a metric that adequately represents the benefits of improving performance across different jobs; and the third is posed by those who maintain that a single measure of general cognitive ability adequately explains the predictability of job performance criteria.

With regard to the use of cost-effectiveness measures, some question the acceptance of such measures because the goal of the armed forces is to win wars, not to make money. Thus, providing estimates of the dollar value of military personnel performance must be reconciled with the *raison d'être* of the military.

We believe cost-benefit analyses are entirely consistent with the underlying objective of winning the war, but at the same time contribute to maximizing productivity gains. Examining the cost-effectiveness of alternative weapons systems is now the routine in making decisions among alternative systems. Similarly, we believe that personnel systems alternatives should and can be subjected to the same type of analysis as weapons systems. Even though the benefits side of utility analyses of personnel systems may rely more on subjective parameter estimates than do hardware systems, utility analyses of personnel systems are not inimical to achieving military objectives.

Increasingly, we see that all components of personnel systems are being subjected to quantitative cost-effectiveness analysis, including recruiting, selection, classification, training, retirement, retention, and promotion. These analyses attempt to arrive at productivity gains considering alternatives in the same manner as is done in the civilian sector.

Such analyses of personnel systems are not surprising because while jobs exist within the military context, they are largely representative of civilian jobs (Hunter, Crosson, and Friedman, 1985). The military job families encompass the spectrum of civilian jobs in the *Dictionary of Occupational Titles*. As has often been stated, the

language of business is dollars, and in recent years studies have been directed towards expressing human resources productivity gains in dollar terms in the civil service and the military.

Estimates of cost-effectiveness obtained through selection and classification is not a new concern (Maier and Fuchs, 1972; Sands, 1973a, 1973b; and Sorenson, 1965). The application of Brogden's utility model of 1949 in the military is a more recent phenomenon (Eaton, Wing, and Mitchell, 1985; Schmidt, Hunter, and Dunn, 1987).

One concern expressed from time to time through the years in the military context is that analyses of the peacetime force may provide results that differ from an analysis of the force in war (i.e., the effective garrison soldier may not be the effective combat soldier). By necessity, most analyses are not analyses of combat. But there is no compelling argument that proficiency and effort are not the best predictors of later performance even in combat. Additionally, the value of that performance also is generally judged in the context of combat scenarios in most simulations.

With regard to the second challenge, some question the assumption that productivity gains can be measured in a common dollar metric across jobs. We employ mean predicted performance (MPP) as the measure of performance in each job and then convert gains in MPP across jobs to arrive at benefits. The concern appears to be that the dollar metric cannot be meaningfully employed across jobs.

We address that concern by focusing on three issues: that utility is a linear function of performance (or predicted performance); that performance differences are of equal importance across jobs; and that the dollar-valued metric is necessary to link costs and benefits.

The basic selection and classification utility equation depends only on linearity. Indeed, behavioral science research in general almost always utilizes this assumption. Hunter and Schmidt (1982) provide a detailed analysis of the question of meeting this statistical assumption. They conclude that "an obsessive concern with statistical assumptions is not justified. This is especially true in light of the fact that for most purposes, there is no need for utility estimates to be accurate down to the last dollar" (p. 245).

The predictor-criterion correlation in utility models uses a proxy criterion, performance, in place of a dollar-valued criterion since the latter is not available for direct measurement. The assumption is then made that both criteria are linearly related to the

predictor and to one another. Schmidt, Hunter, McKenzie, and Muldrow (1979) believe the relationship between the proxy and the dollar-valued criteria to be linear, or if not, the result underestimate utility because ceiling effects lower the correlation between the predictor and the proxy criterion. As noted earlier in this report, it is unlikely that any production function will continue to increase linearly with ever increasing increments of high quality employees--the law of diminishing returns eventually applies. For example, if there are too many high quality employees, they will be assigned to tasks of less value to the organization and thus affect utility estimates. However, in real world situations, with regard to most performance-quality distributions, assuming a linear relationship between performance and value appears to be reasonable.

With regard to our assumption that performance differences are of equal importance across jobs or the output from all jobs is of equal value, this assumption poses no inherent methodological limitation in the use of the MPP measure. By assuming the output from all jobs is of equal value, we accurately reflect the historic reality of manpower policy. Policymakers in all services never have been amenable to directly weighing the relative importance of jobs. The closest they come is to impose minor quality constraints in the assignment system to ensure that the combat arms obtain some of "their fair share."

If policymakers are willing to assign importance value across jobs and/or within a job in the future, MPP can readily be converted to reflect varying job values. Alternatively, rational estimates of SD_y, such as the global estimating procedure, could be used to obtain values separately for each job. As indicated in the manuscript, the type of SD_y estimating procedure employed would not affect the selection of the procedure producing the greatest benefit.

We do not now advocate the conversion of MPP into job measures for importance or value separately for each job because the use of such information in an operational system would require a major policy change. Also, the large research effort required to obtain such important data of separate SD_y estimates for each job would entail resources well beyond the scope of our current project.

We are aware of the large, ongoing Project A effort to obtain ability level/performance values within and across jobs (Nord and White, 1988). We suggest in a subsequent report (Zeidner and Johnson, 1989) the desirability of each service evaluating its productivity gains making their own assumptions and estimates. If this were to be done, Project A job importance values may readily be employed, if desired. Once again,

given the magnitude of differences among procedures considered in the simulation described in Zeidner and Johnson, 1989, it is unlikely that the results obtained would change the recommendations we make. On the other hand, it is likely that different assumptions and parameter estimates would result in different estimates of overall productivity gains. But it is important to note, as we do in the report, that our assumptions and estimates (especially the use of the 40% proportional rule) provide very conservative estimates of dollar gains.

With regard to the issue that an underlying dollar-valued metric is necessary to link costs and benefits, we strongly feel that utility data are necessary if informed military judgment is to be introduced into the assignment process.

Currently, the military services employ a selection and classification system designed in the late 1940s. No advantage is taken of hierarchical classification--neither of disparate criterion means and/or variances across jobs nor of disparate job values. Nor is any advantage taken of computer-based optimum algorithms. Additionally, the ASVAB itself has been allowed to deteriorate as a classification tool because of the almost exclusive focus on raising mean validity and ignoring "differential assignment theory." The nearly fatal blow to classification efficiency is the use of operational aptitude areas in the current system that vitiates the little differential validity still inherent in the ASVAB tests.

We believe that a major reason that these debilitating decisions have been made is the almost exclusive concern of researchers with defining and communicating benefits of selection and classification without an analysis of costs. Our colleagues in the civilian sector now strongly support the approach first proposed by Brogden 40 years ago. We endorse Cascio's (1987a) statement that is particularly germane to this issue of employing dollar values. If we evaluate only in statistical terms, Cascio says, much of what we do in human resources is largely misunderstood and underestimated by the organizations we serve (p. ix). Not only is the language of business dollars, but so too are they the language of government and the military, especially in an era of scarce resources.

The common underlying dollar metric across jobs need never be employed to achieve optimal classification efficiency. But if we leave it at that we are forced to describe gains in terms of validity change only. The current state of affairs in selection and classification, and indeed in human resources utilization as a whole, may be perpetuated indefinitely. By linking costs and benefits, we may be able to change many aspects of the

total personnel system through integrating decisions concerning not only selection and classification, but recruiting, training, and retention as well.

We badly need a mechanism for showing the net benefits of changing selection and classification procedures through changes in the ASVAB measures, increasing recruiting standards, developing and maintaining a computer-based optimal allocation system, and replacing inadequate predictor composites. We believe the best way to do this is to communicate the results of utility analyses in terms of productivity gains in dollars.

The proponents of the third challenge emphasize the power of general cognitive ability in the prediction of real word performance and warn of its implications for both selection and classification procedures. Schmidt, Hunter, and Larson (1988), noting the research of Hunter (1983; 1984; 1985) based on very large military samples, conclude that "general cognitive ability is as good or better predictor of performance in training in most military job families as ability composites derived specifically to predict success in particular job families. These findings are contrary to the current theory that is the foundation of differential assignment of personnel to jobs in the military" (p.1.)

It is important to note explicitly that only differences in predictive validity are considered by Hunter and Schmidt, et al., in reaching their conclusion. We believe if the authors had considered "differential assignment theory" properly, a different conclusion would have been reached.

In differential assignment theory we define the classification efficiency of a set of test composites in terms of the gain in MPP score under optimal assignment conditions over that obtainable using random assignment. The potential classification efficiency of a battery is defined as the gain in the MPP score resulting from optimal assignment that is obtainable using full least squares estimates (FLS) as both assignment and evaluation variables. (See Johnson and Zeidner, 1989, for a detailed discussion of classification efficiency.)

In differential assignment theory any gain or loss in predictive validity is relegated by the underlying mathematics (a result, not an assumption) to a minor role, in most realistic examples, to achieve classification efficiency. If two sets of test composites, neither of which are FLS composites, are compared to measure classification efficiency in terms of MPP, the set of composites showing the smaller average predicted validity could be the most classification efficient. This would result if one set of composites were created to maximize potential validity and the other to maximize classification efficiency.

At the other end of the spectrum, from Hunter and Schmidt, "specific aptitude" theorists pursue the goal of identifying job specific test composites, but also on the basis of predictive validity rather than on the basis of MPP. Pursuing the goal of finding specific composites seriously interferes with the goal of achieving classification efficiency through identification of efficient tests for inclusion in a battery and of forming efficient composites when FLS composites are not used.

The current ASVAB and aptitude area system may be largely attributable to the pursuit of the erroneous specific aptitude theory. The Army's aptitude areas are of questionable value. However, we believe that considerable classification efficiency is potentially obtainable from the existing ASVAB if it is used in accordance with differential assignment theory.

Results of general cognitive ability or specific aptitude theories based on predictive validity tells us very little about either theory as they relate to classification efficiency. The exclusive consideration of predictive validity in evaluating classification efficiency is due to a very common misunderstanding of the psychometrics of classification. The danger of such a misunderstanding by key researchers may eventually destroy the usefulness of the ASVAB for classification purposes to the point where the use of a single selection/ classification measure might be considered by the services.

3. Improved Personnel Utilization Through Classification and Allocation

Cascio (1987b) noted that assigning individuals to independent jobs has any one of three objectives: to assign each individual so that over all individuals, the highest possible predicted performance will result (pure selection); to place each individual on the job for which he or she is best qualified (vocational guidance); or to place persons on jobs so that all jobs are filled by individuals who meet at least some minimum standards of performance (cut and fit). Four types of solutions have been proposed to achieve these objectives: mathematically optimal methods involving differential prediction; methods that classify individuals into jobs according to the discriminant function; methods based on linear programming and goal programming to meet multiple objectives; and non-optimal cut and fit methods (Ghiselli, 1956).

Employee classification in industry is as yet largely unexplored for a variety of reasons, including the complexity of the procedure and the belief that hiring a "generalist" prolongs an organization's return on investment (Zedeck & Cascio, 1984). However, Cascio (1987b) believes that society seems to be espousing more of the classification than

the selection approach to the staffing of organizational roles. In the future world of work, Cascio states, the idea of rejection may have to be abandoned, as society strives to make all individuals productive, and ensure the fullest use of each individual's potential.

In the military context, research continues to address the complex problem of personnel classification through differential prediction modeling and computer-based allocation systems. Military manpower decisionmakers see that one way of improving performance, without increasing costs, is to improve the matching of recruits to jobs, the optimality of matches being judged against an objective function that aggregates performances across jobs.

The matching function is made difficult for a variety of reasons including ever-changing supply and quality mix among entering recruits; changes in requirements or job quotas over time; differences in the criticality or utility of jobs varying as a function of turnover rates; use of low minimum job standards; desire to meet equal opportunity goals; inadequacy of traditional measures of job performance; assignment of individuals to jobs sequentially rather than in batches; and reduced differential validity in the subtests of the Armed Services Vocational Aptitude Battery (ASVAB).

As noted in the introductory chapter, each year the military selects some 315,000 new recruits and decides in which job speciality each new recruit should be trained and assigned. Most of these recruits have little or no civilian work experience; consequently the services rely heavily on educational and aptitude test information.

The ASVAB is the principal means of selection and classification. It is comprised of ten tests, four of which yield the Armed Forces Qualification Test (AFQT) score, a measure of general cognitive ability and trainability. The AFQT score and educational attainment is used to determine enlistment eligibility. Scores on other test composites, called aptitude area composites in the Army, are used to qualify recruits for specific job families.

The Army, like the other services, determines who will be permitted to enlist through the formal and informal enlistment standards it establishes. Formal standards are minimally acceptable composite scores, both for enlistment and job assignment; informal standards are implicit in incentives given to recruiters and in policy guidelines furnished guidance counselors to assist recruits in the determination of their stated job preferences. Fernandez and Garfinkle (1985) state that these formal and informal standards "determine along with the amount of resources devoted to recruiting and the general willingness of

young people to enlist, the quality mix among recruits both for the Army as [a] whole and for individual Army specialities." (p. vi.).

The Army's current operational allocation system used to assign new recruits to MOS considers minimum qualifying aptitude area scores; quality goals in each MOS in order to distribute recruit "quality" more evenly; and the priority of the MOS, including time available to fill its training seats. The present system meets total accession requirements and nearly meets all individual MOS training and quality goal requirements. However, the system has relatively low job-matching efficiency because of low differential validity in ASVAB and because only minimum (low) job standards are considered in making assignments. The average recruit today qualifies for 85% of all MOS in the Army.

A new system is now under development employing a modern person-job match technology that is intended to respond quickly to changing personnel demands, supplies, costs and objectives. This system, called the Enlisted Personnel Allocation System (EPAS), described in more detail in a subsequent report (Zeidner and Johnson, 1989), embraces a multi-stage strategy: a plan is developed for the allocation of recruits expected by the Army over the next year; the plan guides training seat recommendations; and the plan is frequently updated to reflect the most recent supply and demand estimates.

EPAS is planned to extend the process beyond the assignment decision, to include modules that improve supply and demand estimates, and to execute and evaluate decisions. The allocation problem is solved by a network optimization model designed to meet MOS requirements and quality goals, while maximizing performance.

In making efficient assignments using EPAS, the payoff of the best possible assignment of a recruit is compared to payoffs of other possible assignments. Payoff information is used along with other detailed information relating to recruit and job characteristics so that each assignment decision can be evaluated from a payoff prospective.

Fernandez and Garfinkle's (1985) study of setting standards and matching recruits to jobs is particularly important in the context of our empirical study of productivity gains, described in a subsequent report (Zeidner and Johnson, 1989), because of several similarities in the treatment of job performance and costs. To a question of importance to military manpower policymakers: Is there any objective basis for setting standards for enlistment, either into a service as a whole or into specific jobs, or for determining the "right" job for each recruit? Fernandez and Garfinkle's answer was a qualified yes.

They used a version of Armor, Fernandez, Bers, and Schwarzbach's (1982) cost/performance tradeoff model for examining optimal enlistment standard units. The authors examined optimal recruit assignment by using a performance measure developed for four Army jobs.

Three concepts in their study are highlighted here because they are also used in our study of classification utility. First, their study provides an analysis of recruiting and force costs from an economic perspective. Fernandez and Garfinkle pointed out that raising the standard for a job increases its recruitment of costly-to-recruit "high-quality" enlistees (high school graduates who score above the 50th percentile on the AFQT). The Army obviously would prefer the highest quality recruit possible, all other things being equal. But the Army cannot, of course, fill all of its jobs with high-quality recruits (even if enough of them reached enlistment age each year), because to do so it would have to outbid the other military services and all potential civilian employers. This would be prohibitively expensive (even turning to the draft).

Although attracting more high-quality recruits requires expending more resources, costs must be balanced against the improved performance of high-quality recruits. This economic concept of cost/performance tradeoff as it relates to recruiting and force costs is one that is used in our study.

Second, Fernandez and Garfinkle use a summary measure, qualified man-months (QMM), developed by Armor et al. (1982). A qualified man-month is defined as a work (i.e., post-training) month contributed by an enlistee who is able to perform his or her job at least at the minimum acceptable level of competence. A version of QMM called "productive man-months" (PMM), is used as a common baseline figure in our study detailed in a subsequent report (Johnson and Zeidner, 1989). In this context, we use PMM as an alternative to employee flows models.

Third, the authors' study employs, as one criterion measure, attrition. Remaining in the service is an important component of an enlistee's job performance; it should be considered in evaluating the relative cost-effectiveness of various groups of enlistees. Training costs, including the enlistee's pay during this non-productive period, constitutes a large portion of the total cost of maintaining an individual through the first tour of duty, typically three years. Much of the attrition occurs during the first five months (marking the end of advanced training), particularly during the first three months (during basic training).

Fernandez and Garfinkle suggested that an enlistee who is barely competent but completes the initial tour of duty might be worth more to the Army than an enlistee who performs perfectly for several months and leaves. This concept of differential payoff as a function of tenure is also considered in our study.

In summary, the authors found the following: high AFQT scorers outperform low scorers by a considerable margin, attributable in part to reduced attrition rates of high AFQT scorers; a smaller force can produce as many units of performance (QMM) as a larger force, and at less cost; using higher standards than the Army is currently using would appear to be justified, but giving a firm answer depends on how much is better performance worth and to what extent is performance more important in one job than another; and the QMM is judged to be a useful means of combining two very different measures of individual job performances, quality of performance and availability to perform. As the reader will see, these findings are consistent with our utility findings of selection and classification standards described in a subsequent report (Zeidner and Johnson, 1989).

D. CREDIBILITY AND CURRENT SELECTION DECISIONS

We have noted that improving productivity has been a subject of major concern during the last decade. Labor and management are now actively working for ways of improving productivity. Everyone acknowledges that people are the key to productivity, and that productivity gains depend greatly on matching the attributes of people with the demands of the job.

From the time of the Army's success with the Alpha tests during World War I, employers were eager to capitalize on the use of standardized tests in the hiring process. The tests were shown to be valid predictors of job performance and perceived as being fair. But to assess the practical impact of findings, practitioners resorted to the use of difficult-to-understand statistical concepts. Starting with the first decision theoretic models, a new language of translating validity findings began to emerge, and eventually it became possible to make economically meaningful "bottom-line" statements.

In recent years, decision models have become even more realistic, comprehensive, integrative and accurate. They permit comparisons to be made among alternative investment strategies on the same basis as other organizational decisions.

In this section, the last dealing with selection utility, it may be useful to note briefly milestones in the history of the development of decision theoretic utility analysis models.

The Taylor-Russell (1939) model reformulated the concept of validity away from individual prediction to institutional decisionmaking and redefined the concept of measurement accuracy to that of accuracy of predicting decision outcomes. Their model was the first to show that the context of a selection decision must be considered to evaluate its value.

Brogden (1946a) showed that the validity coefficient itself is a direct index of selective efficiency, (e.g., a predictor with a validity of 0.50 would produce 50% of the gain resulting from using a perfect selection device). Brogden's (1949) utility formulation, the model that is the basis of all later elaborations, was the first to consider payoff in dollar terms, costs and other external parameters of the selection situations.

Cronbach and Gleser (1957, 1965) in their influential book, *Psychological Tests and Personnel Decision*, firmly established decision theory as the appropriate framework for developing and applying tests. They demonstrated that every decision problem must be specified and these specifications must be used to determine the appropriate mathematical model.

Schmidt, Hunter and co-workers (1979) developed a practical procedure for estimating SD_y that could replace cost accounting procedures. Using this rational method, along with validity generalization findings, they demonstrated, in a series of realistic utility studies, that productivity gains attributable to selection were very large--in the millions of dollars per year. In 1982, they generalized the basic utility model, making it applicable to all personnel interventions intended to improve job performance. In 1983, they cumulated data showing that the standard deviation of employee output can be conservatively estimated at 20% of mean output and 40% of mean salary.

Boudreau (1983) enlarged basic utility formulations by incorporating economic considerations--variable costs, taxes, and discounting. He redefined the payoff function to better reflect economic concepts used in organizational investment decisions. Also in 1983, he extended previous utility models by incorporating the flow of employees into and out of the work force (and in 1985, with Berger, developed a more general external employee movement model). In 1984, Boudreau applied break-even analysis as a means of simplifying decisions. In 1987, Rich and Boudreau further enhanced utility models by incorporating risk assessment techniques.

During the 1980s, a number of investigations developed SD_y estimating techniques to increase its reliability, understandability and credibility (Bobko et al., 1983; Burke & Frederick, 1986; Cascio & Ramos, 1986; Eaton et al., 1985). Greer and Cascio (1987), in a significant study, found that the external validity of behaviorally based SD_y estimates were quite accurate when compared to a carefully developed cost accounting procedure. They also found that behaviorally based estimates were credible to managers.

The cumulative findings of utility analysis research strongly suggest that it will improve organizational decisions. But it is curious to note that nearly all utility "applications", published or unpublished, were either demonstrations of large potential productivity gains of programs if they were to be adopted by organizations, illustrations of the uses and advantages of new elaborations of the basic utility model, or justifications of use of existing personnel programs in organizations.

What appears to be much needed are data on "real" applications of utility analysis that are intended for use in the process of organizational decisionmaking. Real applications would accelerate the development of theory and technology of utility analysis in personnel interventions.

The reader may raise the question: Why have so few utility analysis applications been used in actual, real-world decision situations, despite the availability of realistic, comprehensive models? We offer a number of possible reasons including the following: psychologists' historically greater interest in effectiveness than in cost-effectiveness; the limited appreciation of psychologists and managers of the great economic impact of selection programs; the lack of widespread understanding of utility analysis concepts and formulations; lingering concerns about required statistical assumptions in utility analysis and of assumptions and estimations needed for determining utilities in specific decision situations; and the reduced opportunities afforded investigators to participate in the development and evaluation of selection program decisions.

The last point is especially significant in understanding the role of testing in the context of equal employment opportunity considerations during the last two decades. It is widely acknowledged that employers, both in the private and civil service sectors, have been extremely cautious in using tests in hiring decisions. Such caution is attributable in part to the perceived negative impact of selection tests on the lives of minorities; to achieve more balance in minority representation in the work force; and to avoid possible litigation.

Employer concerns are directly reflected in the use of the Professional and Administrative Career Examination (PACE) in hiring civil servants. As noted earlier, on the basis of a suit brought against the Office of Personnel Management (OPM) on the grounds that PACE had an adverse impact on minorities, the government entered into a consent decree agreeing to eliminate the PACE. The PACE had been administered to nearly 200,000 applicants yearly who were interested in obtaining positions in more than 100 entry-level jobs. No objective tests were administered by OPM in hiring for these jobs during the six years subsequent to the consent decree.

In June of 1988, the OPM announced a proposal for a revised procedure. Judith Havemann (The Washington Post, June 23, 1988) reported:

> Six years after the federal government's entrance test for workers was thrown out as racially discriminating . . . it intends to abandon the traditional written examination requirements for civil servants and allow college graduates with top grades to be hired on the spot.
>
> Applicants for entry-level professional and administrative jobs will become eligible for hiring by either earning a high college grade point average or passing a job-related skills test and a wholly new type of test called an Individual Achievement Record [biodata] which attempts to measure "the full range of relevant personal qualities required for successful job performance" according to the Office of Personnel Management.
>
> The government has no entrance examination for more than 100 jobs since the Professional and Administrative Examination was thrown out in 1982. Hiring has been based on interviews, recommendations and college grades.
>
> . . . Minorities do far better on the Individual Achievement Record than on written ability tests, in comparison with whites, according to a study of 6,000 federal workers. Although some discrepancy remains between average black and white scores on the Individual Achievement Record, it is less than one-quarter as much as on traditional tests such as PACE. . . . OPM said that the Individual Achievement Record predicts successful on-the-job performance to a high degree.
>
> Donald J. Devine [OPM's Director Constance Horner's predecessor] called her proposal "a sad day for the civil service when it can't have an objective civil service exam, but I can't criticize OPM because I know the kind of pressure they're under from the law suit." (p. 1,9)

According to Judith Havemann's Washington Post report of the next day, the National Treasury Employees Union filed suit to block the government's proposed new hiring system saying that it was illegal and irrational to hire college students based on their grade point averages without subjecting them to traditional examinations. According to the union, "the law says there shall be nationwide examinations administered twice a year."

Constance Horner (the OPM Director) expressed perplexity that a proposal would be grounds for a suit, "especially a proposal arrived at a broad consensus that holds such hope for bringing together the values of merit and equity." Clarence Thomas, Chairman of the Equal Opportunity Commission, called Horner's plan "absolutely fantastic," and helpful to minorities and women.

These opposing views on OPM's proposal were aptly embraced earlier in Haney's (1981) conclusion that the role of standardized testing is both advocated and challenged in technical terms, but the prominent social concerns surrounding testing are rooted in matters of social and political values.

Within the context of utility analysis, it would be interesting to examine the technical issues concerning PACE and its proposed replacement with regard to productivity and testing "fairness".

Three validity studies pertain: Schmidt et al. (1986) studied the productivity of the federal work force with the use of PACE; Hunter and Hunter's (1984) meta-analysis provided validity data for various predictor types suitable for entry-level jobs against the criterion of supervisory rating; and McHenry (1987) combined various predictor types and alternatives against separate criterion dimensions within a job.

Schmidt et al. (1986) determined a generalized validity for the PACE of 0.56 for white-collar civil service jobs within the middle range of job complexity, the relevant job complexity range for PACE use.

Hunter and Hunter (1984) found the validity of general mental ability tests (similar to the PACE) was 0.53; for biographical inventories (or biodata), 0.37; and for academic achievement, 0.11. Skill tests (similar to OPM's Individual Achievement Record) were considered by them as not suitable for use with entry-level jobs, but we estimate that their validity would have been low in any case. Also, the combined validity, given the meta-analytic results for all three measures proposed by OPM, if appropriately weighted, would not have exceeded .40.

McHenry (1987) reported a validity of 0.63 for a general ability composite (similar to PACE) against a well-developed job proficiency criterion, and a validity of 0.31 for the same composite against a motivationally based supervisory ratings criterion. For biodata the validities were 0.26 and 0.33, respectively. However, for both test types considered together against a combined criterion, weighted by importance, we estimate the validity to be around 0.65.

Using the data of all three studies, then, we estimate the potential validity of an expanded PACE-like exam to be about 0.65 and the potential validity of OPM's proposal to be about 0.40. However, for computation of potential productivity gains, we compare the validity of 0.56 for the PACE against the validity of 0.40 for OPM's proposal. Remembering that validity enters into the utility model as a multiplicative factor [see Equation (2.6), for example], the potential loss in overall utility of reducing validity by 28% (0.56 - 0.40) is large.

Adjusting Schmidt et al.'s (1986) results to pertain only to GS 5-7 grade levels, the entry-level grades of concern, the potential productivity gain from one year's use of the PACE is $1.71 billion and from ten year's use, $17.20 billion. If OPM's proposal were to be used as the basis of new hires, the potential gain would be reduced by $0.48 billion for one year's use and by $4.82 billion for ten year's use.

Turning to the issue of testing fairness, The National Academy of Sciences Committee on Ability Testing defines ability tests as measures of *developed* abilities; they serve as indicators of ability to learn (Wigdor & Garner, 1982). General ability measures reflect the broad range of knowledge that results from growing up and attending school in a twentieth century English culture (Anastasi, 1984). Mental ability test scores are subject to improvement as educational experience, both in and out of school, causes these abilities to develop; ability tests are not measures of some inborn and unchanging capacity.

Ability tests have proved valid in all situations and for all jobs (Schmidt & Hunter, 1981). No subgroup parity model meets both the goal of selecting "the most qualified" and the goal of equality of selection outcomes to achieve selection "fairness." This is so because mean ability test scores of some minorities are lower than the mean ability scores of the majority.

In the 1970s, a concept of selection "fairness" emerged with the view of promoting minority representation in the work force. Fairness advocates called for compensatory hiring to maintain equality of selection outcomes for various subgroups. Some proponents of "fairness" even argued for the elimination of ability tests (e.g., since the PACE exam screens out about 95% of black applicants, its use should be abandoned).

The most commonly accepted model of test fairness is the regression model (Cleary & Hilton, 1968). This model defines a test as unfair to a minority group if it predicts lower levels of job performance than the group actually achieves, a concept incorporated in the Uniform Guidelines on Employment Selection Procedures (EEOC, 1978).

The accumulated evidence of the fairness of tests is supported by numerous studies. Tests predict the job performance of a minority and the majority in the same way. Lower test scores among minorities are accompanied by lower job performance, exactly as in the case of the majority (Schmidt & Hunter, 1981). Schmidt and Hunter assert that their research findings show that employment tests do not cause "adverse impact" and that differences between groups are directly reflected in job performance and thus are real. They are not created by the tests.

Bersoff (1984) pointed out that it was the Chinese over 3,000 years ago, not the Americans in this century, who first used large-scale testing (Dubois, 1966); but it was in the United States that the method was enthusiastically supported. It is Bersoff's contention that what appears to be an anti-testing movement in this country and in the Congress is not an anti-test movement at all. He suggested that the law's concern has been evoked by three major social developments: society is attempting to undo the effects of history of *de jure* segregation and discrimination against racial and ethnic minorities; recognition by the courts as a constitutional imperative, of the right against impermissible intrusion by the government into the private lives of its citizens (e.g., as may be the case in the use of personality and attitude tests); and stupidity in the failure to use reasonable care in carrying out one's obligations (e.g., faulting both psychologists and judges for increased regulation of testing).

Bersoff asserts that if psychologists are to be respected by the courts and treated as more than mere numerologists, they must offer situation-specific, ecologically valid, objective data that serve science, not a particular adversary.

In that light, it appears to us that the examinations of OPM's proposed revision in federal hiring, viewed in terms of its potential economic impact on productivity, may contribute to decisions that can better withstand societal, judicial and scientific scrutiny.

We turn in the subsequent reports to issues of classification efficiency (Johnson and Zeidner, 1989) and classification utility (Zeidner and Johnson, 1989) in the military. As noted earlier, the ASVAB is the principal means of selecting and classifying over 300,000 new recruits each year. Most of these recruits have little or no civilian work experience; the services rely heavily on educational and aptitude test information. It is of interest that service practices are consistent with the Uniform Guidelines, but are not legally constrained by the Guidelines. It is in the military context, then, that we can best observe the full

impact of decision theoretic approaches on personnel selection and classification systems in operations.

GLOSSARY

ability test[a]--A test that measures the current performance or estimates future performance of a person in some defined domain of cognitive, psychomotor, or physical functioning.

achievement test[a]--A test that measures the extent to which a person commands a certain body of information or possesses a certain skill, usually in a field where training or instruction has been received.

adaptive testing[a]--A sequential form of testing in which successive items in the test are chosen based on the responses to previous items.

algebraic variability derivation--A technique for incorporating uncertainty into utility by the use of variance estimates.

allocation efficiency--The gain in benefit over random assignment obtained from an optimal assignment process attributable to differential validity.

allocation process--Classification that capitalizes on differential job validity.

alternative[c]--A course of action whose selection may result in an outcome that will attain the original objective.

aptitude test[a]--A test that estimates future performance on other tasks not necessarily having evident similarity to the test tasks. Aptitude tests are often aimed at indicating an individual's readiness to learn or to develop proficiency in some particular area if education or training is provided. Aptitude tests sometimes do not differ in form or substance from achievement tests, but may differ in use and interpretation.

assessment procedure[a]--Any method used to measure characteristics of people, programs, or objects.

attenuation[a]--The reduction of a correlation or regression coefficient from its theoretical true value due to the imperfect reliability of one or both measures entering into the relationship.

battery[a]--A set of tests standardized on the same population, so that norm-referenced scores on the several tests can be compared or used in combination for decision making.

behavior[b]--Observable aspects of a person's activities.

benefit--A theoretically desirable measure of performance that is value-weighted for jobs and validity in terms of an appropriate metric; when the benefit measure is correctly combined with costs, it provides a measure of utility.

break-even values--The determination of the lowest value of any individual parameter that would still yield a positive total utility value.

classification--The matching of individuals and jobs in an organization with the goal of maximizing aggregate performance; it requires multiple predictors jointly measuring more than one dimension and multidimensional job criteria.

classification[a]--The act of determining which of several possible job assignments a person is to receive.

classification battery-- A battery of tests used operationally to classify personnel.

classification efficiency--The gain in benefits over random assignment obtained from an optimal assignment process attributable to allocation and hierarchical classification efficiency; a separate LSE must be used for each criterion.

cognition[c]--The act or process of knowing, including both awareness and judgment.

composite score[a]--A score that combines several scores by a specified formula.

concurrent criterion-related validity[a]--Evidence of criterion-related validity in which predictor and criterion information are obtained at approximately the same time.

construct[a]--A psychological characteristic (e.g., numerical ability, spatial ability, introversion, anxiety) considered to vary or differ across individuals. A construct (sometimes called a latent variable) is not directly observable; rather it is a theoretical concept derived from research and other experience that has been constructed to explain observable behavior patterns. When test scores are interpreted by using a construct, the scores are placed in a conceptual framework.

cost accounting approach--The approach used to develop a dollar criterion that considers the value of products and services and the organization's costs to provide products and services.

cost effectiveness[c]--A state or condition in which the benefits associated with a particular outcome clearly exceed the cost of obtaining the outcome.

decision[c]--A moment of choice in an ongoing process of evaluating alternatives with a view to selecting one or some combination of them to attain the desired end.

decision tree[c]--A framework for developing the anatomy of a decision making situation that uses the concepts of probability, utility, and expected value.

decision theoretic approach--The set of alternatives, costs and possible outcomes leading to a choice.

differential validity--The level of prediction using LSEs of differences among criterion scores when referring to H_d; this measure is related to the variation of a validity vector with jobs and to an assignment variable being more valid for its own job family than any other job family.

discounting--A procedure for equating the costs and benefits that accrue over time to reflect the opportunity costs and returns foregone.

efficiency--A solution that minimizes costs as measured by physical resources and time utilized.

expected value[c]--A concept that permits a decision maker to place a monetary or other value on the positive and negative consequences likely to result from the selection of a particular alternative.

external employee movement--The analysis of employee separations and acquisitions in an organization.

goal[c]--A subset of an objective expressed in terms of one or more specific dimensions.

gross national product--The sum of all expenditures on goods and services by households, by firms on new capital, and by government.

hierarchical classification efficiency--All classification efficiency not explainable as allocation efficiency; it capitalizes on disparate means and variances of the benefit scores for the corresponding jobs.

hierarchical layering--A phenomenon in which LSEs are more valid or of more value for some jobs than for others.

human capital--The skills of the workforce that determine what workers can contribute to the production process.

human resource accounting--The economic consequences of employees' behavior.

inter-rater reliability[a]--Consistency of judgments made about people or objects among raters or sets of raters.

interest inventory[a]--A set of questions or statements that is used to infer the interests, preferences, likes, and dislikes of a respondent.

inventory[a]--A questionnaire or checklist, usually in the form of a self-report, that elicits information about an individual. Inventories are not tests in the strict sense; they are most often concerned with personality characteristics, interests, attitudes, preferences, personal problems, motivation, and so forth.

item analysis[a]--The process of assessing certain characteristics of test items, usually the difficulty value, the discriminating power, and sometimes the correlation with an external criterion.

job analysis[a]--Any of several methods of identifying the tasks performed on a job or the knowledge, skills, and abilities required to perform that job.

job relatedness[b]--The inference that scores on a selection instrument are relevant to performance or other behavior on the job; job relatedness may be demonstrated by appropriate criterion-related validity coefficients or by gathering evidence of the relevance of the content of the selection instrument, or of the construct measured.

joint probability[c]--The probability that two or more events will occur.

labor--The worker effort available to the production process.

law of diminishing returns--As the quantity of an input is increased and the quantity of other inputs stays the same, a point is reached where the additional output produced per unit of added input declines.

linear combination[b]--The sum of scores, whether weighted differentially or not, on different assessments to form a single composite score.

linear model[c]--A model of choice in which the evaluation of each alternative is based on the sum of its weighted values on all its dimensions, and the alternative with the greatest sum is the obvious choice.

longitudinal study[a]--Research that involves the measurement of a single sample at several different points in time.

marginal cost--The cost of producing an additional unit.

maximizing behavior[c]--An approach to decision making oriented toward obtaining an outcome of the highest quantity or value.

mean predicted performance (MPP)--The measurement of benefits can be approximated by computing MPP across jobs; if MPP is weighted by the value of each job, it becomes a more useful measure of benefits. It provides a means of comparing the effectiveness of alternative tests or test batteries in the context of a specified set of jobs and performance scores.

meta-analysis[b]--A procedure to cumulate findings from a number of validity studies to estimate the validity of the procedure for the kinds of jobs or groups of jobs and settings included in the studies.

meta-analysis--A technique for determining the degree to which the variance in validity coefficients across situations for job-test combinations is due to statistical artifacts.

model[c]--A physical or abstract representation of some part of the real world that is used to describe, explain, or predict behavior.

Monte Carlo analysis--A stochastic technique that can provide numerical solutions for mathematical functions lacking analytic solutions; the analysis typically uses random numbers as input to an evaluation process employing variance reduction procedures.

multidimensional screening (MDS)--A selection/classification process using an algorithm that insures no nonselected person has a higher predicted performance on any job than the person assigned to that job; the algorithm also ensures that no other assignment can further raise the mean predicted performance.

multivariate[b]--Characterizing a measure or study that incorporates several variables.

norms[a]--Statistics or tabular data that summarize the test performance of specified groups, such as test takers of various ages or grades. Norms are often assumed to represent some larger population, such as test takers throughout the country.

norm-referenced test[a]--An instrument for which interpretation is based on the comparison of a test taker's performance to the performance of other people in a specified group.

objective[b]--Pertaining to scores obtained in a way that minimizes bias or error due to different observers or scores.

operational efficiency--The improvement in MPP obtained from the usually imperfect operational selection assignment process as contrasted to potential efficiency, the improvement obtainable if the maximally efficient prediction composites of a given battery were to be used in optimal selection/assignment algorithms.

opportunity cost[c]--The cost of the next best alternative that is sacrificed to select what appears to be the best alternative.

payoff[c]--The intersection of an alternative and a state of nature in a payoff table; it measures the value (utility) to the decision maker likely to result from the selection of that alternative given the probabilistic occurrence of the state of nature.

payoff table[c]--A convenient framework in which to present the elements of a decision making situation employing the concepts of probability, utility, and expected value.

percentile[a]--The score on a test below which a given percentage of scores fall.

performance[b]--The effectiveness and value of work behavior and its outcomes.

personality inventory[a]--An inventory that measures one or more characteristics that are regarded generally as psychological attributes or interpersonal skills.

placement--A procedure in which individuals are matched to levels within jobs as contrasted to the classification process of matching personnel to jobs.

potential allocation efficiency--The maximum allocation effectiveness achievable from the differential validity of a given test battery and set of jobs expressed as a mean predicted performance standard score.

potential classification efficiency--The maximum classification effectiveness achievable from a given test battery and set of jobs expressed as a mean predicted performance standard score; it incorporates both potential allocation efficiency and hierarchical layering effects.

potential selection efficiency--Rank-ordering applicants on some benefit continuum and rejecting all those below some point on that continuum.

potential utilization efficiency--The sum of potential selection efficiency and potential classification efficiency.

predictive criterion-related validity[a]--Evidence of criterion-related validity in which criterion scores are observed at a later date, for example, for job or school performance.

predictor[a]--A measurable characteristic that predicts criterion performance such as scores on a test, evidence of previous performance, and judgments of interviewers, panels, or raters.

productivity--The ratio of outputs to inputs of a resource (workers, capital equipment); a measure of the degree of the use of resources..

psychometric[a]--Pertaining to the measurement of psychological characteristics such as abilities, aptitudes, achievement, personality, traits, skill, and knowledge.

regression equation[b]--An algebraic equation used to predict criterion performance from predictor scores.

relevance[b]--The extent to which a criterion measure reflects important job performance dimensions or behaviors.

reliability[a]--The degree to which test scores are consistent, dependable, or repeatable, that is, the degree to which they are free of errors of measurement.

reliability coefficient[a]--The square of the correlation of an observed score with its "true" component; often measured as the coefficient of correlation between two administrations of a test. The conditions of administration may involve variation of test forms, raters or scorers, or passage of time. These and other changes in conditions give rise to qualifying adjectives being used to describe the particular coefficient, e.g., parallel form reliability, rater reliability, test retest reliability, etc.

residual score[a]--The difference between the observed and the true or predicted score.

restriction of range[a]--A situation in which, because of sampling restrictions, the variability of data in the sample is less than the variability in the population of interest.

risk[c]--A common state or condition in decision making characterized by the possession of incomplete information regarding a probabilistic outcome.

sample[b]--The individuals who are actually tested from among those in the population to which the procedure is to be applied.

score[a]--Any specific number resulting from the assessment of an individual; a generic term applied for convenience to such diverse measures as test scores, estimates of latent variables, production counts, absence records, course grades, ratings, and so forth.

selection--A procedure for rejecting some applicants for organizational membership as contrasted to assigning all applicants to jobs (classification); or rejecting an applicant for a single job as contrasted to selection and assignment to one of a number of jobs (multidimensional selection).

selection decision[a]--A decision to accept or reject applicants for a job on the basis of information.

selection instrument[b]--Any method or device used to evaluate characteristics of persons as a basis for accepting or rejecting applicants.

selection procedures[b]--Process of arriving at a selection decision.

sensitivity analysis--An analytic technique in which a utility parameter is varied through a range of values, holding other parameter values constant to determine the impact on the total utility estimates.

shrinkage[a]--Refers to the fact that a prediction equation based on a first sample will tend not to fit a second so well.

shrinkage correction[b]--Adjustment to the multiple correlation coefficient for the fact that the beta weights in a prediction equation cannot be expected to fit a second sample as well as the original.

simulation model[c]--A special type of abstract model that is analogous to a segment of the real world and contains a time dimension. It is used to explain and predict behavior as if it occurred in the real world.

skill[b]--Competence to perform the work required by the job.

split-half reliability coefficient[a]--An internal analysis coefficient obtained by using half the items on the test to yield one score and the other half of the items to yield a second, independent score. The correlation between the scores on these two half-tests, stepped up via the Spearman-Brown Formula, provides an estimate of the alternate-form reliability of the total test.

standard score[a]--A score that describes the location of a person's score within a set of scores in terms of its distance from the mean in standard deviation units.

standardized prediction[b]--A test employed for estimating a criterion of job performance, the test having been developed and normative information produced according to professionally prescribed methods as described in standard reference works.

standards[c]--Criteria against which the results of an implemented decision can be measured.

state of nature[c]--A state or condition likely to prevail when a choice is made.

sunk costs--Costs that once incurred cannot be changed by future action.

test[b]--A measure based on a sample of behavior.

test fairness--The most commonly accepted model of test fairness is the regression model; a fair test predicts the job performance of a minority and the majority in the same way.

test-retest coefficient[a]--A reliability coefficient obtained by administering the same test a second time to the same group after a time interval and correlating the two sets of scores.

trade-off value[c]--A value that exists when a given amount of one kind of performance may in some measure be substituted for another kind of performance.

traditional selection approach--The view of tests as measuring instruments intended to assign accurate values to attributes of an individual stressing precision of measurement and estimation rather than selection outcomes.

unidimensionality[a]--A characteristic of a test that measures only one latent variable.

utility[c]--Technically, want-satisfying power; it is often defined as the preference of the decision maker for a given outcome.

utility analysis--The determination of institutional gain or loss (outcomes) anticipated from various courses of action usually measured in terms of dollars.

validity[a]--The degree to which a certain inference from a test is appropriate or meaningful.

validity coefficient[a]--A coefficient of correlation that shows the strength of the relation between predictor and criterion.

validity generalization[a]--Applying validity evidence obtained in one or more situations to other similar situations on the basis of simultaneous estimation, meta-analysis, or synthetic validation arguments.

values[c]--The nominative standards by which human beings and organizations are influenced in their choices.

variability[b]--The spread or scatter of scores.

variable[a]--A quantity that may take on any one of a specified set of values.

variance[a]--A measure of variability; the average squared deviation from the mean; the square of the standard deviation; and, in the experimental design literature, the sum of the squared deviation from its mean doubled by the degrees of freedom.

Z-score[a]--A type of standard score scale in which the mean equals zero and the standard deviation equals one unit for the group used in defining the scale.

NOTES:

a Adapted from American Psychological Association, American Educational Research Association, and National Council on Measurement in Education (1985). *Standards for Education and Psychological Testing*.

b Adapted from Society for Industrial and Organization Psychology (1987). *Principles for the Validation and Use of Personnel Selection Procedures*.

c Adapted from Heyne (1988). *Microeconomics*.

REFERENCES

Abelson, M.A., and Baysinger, B.D. (1984). "Optimal and dysfunctional turnover: Toward an organizational level model." *Academy of Management Review*, **9**, 331-341.

Alexander, R.A., and Barrick, M.R. (1986). *Estimating the standard error of projected dollar gains in utility analysis*. Paper presented at the first annual meeting of the Society of Industrial and Organizational Psychologists, Chicago.

Alexander, R.A., Barrett, G.U., and Doverspike, D. (1983). An explication of the selection ratio and its relationship to hiring rate. *Journal of Applied Psychology*, **68**, 342-344.

American Psychological Association, American Educational Research Association, and National Council on Measurement in Education (1985). *Standards for educational and psychological tests*. Washington, DC: American Psychological Assocation.

Anastasi, A. (1984). Aptitude and achievement tests: The curious case of the indestructible strawperson. In B.S. Plake (Ed.), *Social and technical issues in testing. Implications for test construction and usage*. (pp. 129-140). Hillsdale, NJ: Lawrence Erlbaum.

Anastasi, A. (1988). *Psychological Testing* (6th ed.). New York: Macmillan.

Armor, D.J., Fernandez, R., Bers, R., and Schwarzbach, D. (1982, September). *Recruit aptitudes and Army job performance: Setting enlistment standards for infantrymen*. R-2874-MRAL. Santa Monica, CA: The Rand Corporation.

Arnold, J.D., Rauschenberger, J.M., Soubel, W.G., and Guion, R.M. (1982). *Journal of Applied Psychology*, **5**, 588-604.

Barnes, R.M. (1937). *Time and motion study*. New York: Wiley.

Barnes, R.M. (1958). *Time and motion study* (4th ed.). New York: Wiley.

Bartlett, C.J., Bobko, P., Mosier, S.B., and Hannan, R. (1978). Testing for fairness with a moderated multiple regression strategy: An alternative to differential analysis. *Personnel Psychology*, **31**, 233-241.

Becker, G.S. (1964). *Human Capital*, New York: National Bureau of Economic Research.

Belgrave, F., and Nogami, G. (1986, August). *Variables related to attrition in the Army*. Paper presented at the meeting of the Association of Black Psychologists Convention, Oakland, CA.

Bersoff, D.N. (1984). Social and legal influences on test development and usage. In B.S. Plake (Ed.), *Social and technical issues in testing. Implications for test construction and usage.* (pp. 87-109). Hillsdale, NJ: Lawrence Erlbaum.

Bierman, H., Bonini, C.P., and Hausman, W.H. (1981). *Quantitative analysis for business decisions.* Homewood, IL: Irwin.

Black, M. (1988). Job performance and military enlistment standards. In B.F. Green, H. Wing, and A.K. Wigdor (Eds.), *Linking military enlistment standards to job performance.* Committee on the Performance of Military Personnel Commission on Behavioral and Social Sciences and Education, National Research Council. Washington, DC: National Academy Press.

Bobko, P., Karren, R., and Parkington, J.J. (1983). Estimation of standard deviations in utility analyses: An empirical test. *Journal of Applied Psychology*, **68**, 170-176.

Boehm, V.R. (1977). Differential prediction: A methodological artifact? *Journal of Applied Psychology*, **62**, 146-154.

Boudreau, J.W. (1983a). Economic considerations in estimating the utility of human resource productivity improvement programs. *Personnel Psychology*, **36**, 551-576.

Boudreau, J.W. (1983b). Effects of employee flows on utility analyses of human resource productivity improvement programs. *Journal of Applied Psychology*, **68**, 396-406.

Boudreau, J.W. (1984). Decision theory contributions to HRM research and practice. *Industrial Relations*, **23**, 198-217.

Boudreau, J.W. (December 1988). *Utility analysis for decisions in human resource management.* Working Paper 88-21. Center for Advanced Human Resources Study, Ithaca, NY: NY SSILR--Cornell University.

Boudreau, J.W., and Berger, C.J. (1985). Decision-theoretic utility analysis applied to external employee movement. *Journal of Applied Psychology* [Monograph], **70**, 581-612.

Boudreau, J.W., and Rynes, S.L. (1985). The role of recruitment in staffing utility analysis. *Journal of Applied Psychology*, **70**, 354-366.

Brealey, R., and Myers, S. (1984). *Principles of corporate finance.* New York: McGraw Hill.

Brogden, H.E. (1946a). On the interpretation of the correlation coefficient as a measure of predictive efficiency. *Journal of Educational Psychology*, **37**, 65-76.

Brogden, H.E. (1946b). An approach to the problem of differential prediction. *Psychometrika*, **11**, 139-154.

Brogden, H.E. (1949). When testing pays off. *Personnel Psychology*, **2**, 171-183.

Brogden, H.E. (1951). Increased efficiency of selection resulting from replacement of a single predictor with several differential predictors. *Educational and Psychological Measurement*, **11**, 173-196.

Brogden, H.E. (1954). A simple proof of a personnel classification theorem. *Psychometrika*, **19**, 205-208.

Brogden, H.E. (1955). Least squares estimates and optimal classification. *Psychometrika*, **20**, 249-252.

Brogden, H.E. (1959). Efficiency of classification as a function of number of jobs, percent rejected, and the validity and intercorrelation of job performance estimates. *Educational and Psychological Measurement*, **19**, 181-190.

Brogden. H.E., and Taylor, E.K. (1950) The dollar criterion--applying the cost accounting concept to criterion construction. *Personnel Psychology*, **3**, 133-154.

Burke, M.J., and Day, R.R. (1986). A cumulative study of the effectiveness of managerial training. *Journal of Applied Psychology*, **71**, 232-245.

Burke, M.J., and Frederick, J.T. (1984). Two modified procedures for estimating standard deviations in utility analyses. *Journal of Applied Psychology*, **69**, 482-489.

Burke, M.J., and Frederick, J.T. (1986). A comparison of utility estimates for alternative SD_y estimation procedure. *Journal of Applied Psychology*, **71**, 334-339.

Campbell, A.K. (1979). Statement for Hearing on the PACE before the Subcommittee on Civil Service, Committee on Post Office and Civil Service, U.S. House of Representatives, May 15, Washington, DC.

Campbell, J.P. (1986, August). *Project A: When the textbook goes operational.* Paper presented at The American Psychological Association Annual Meeting, Washington, DC.

Cascio, W.F. (1980). Responding to the demand for accountability: A critical analysis of three utility models. *Organizational Behavior and Human Performance*, **25**, 32-45.

Cascio, W.F. (1982). *Costing human resources: The financial impact of behavior in organizations.* Boston: Kent Publishing Co.

Cascio, W.F. (1987a). *Costing human resources: The financial impact of behavior in organizations* (2nd ed.). Boston: PWS-Kent.

Cascio, W.F. (1987b). *Applied psychology in personnel management* (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Cascio, W.F., and Ramos, R. (1986). Development and application of a new method for assessing job performance in behavioral/economic terms. *Journal of Applied Psychology*, **1**, 20-28.

Cascio, W.F., and Silbey, V. (1979). Utility of the assessment center as a selection device. *Journal of Applied Psychology*, **64**, 107-118.

Cattin, P. (1980). Estimation of the predictive power of a regression model. *Journal of Applied Psychology*, **65**, 407-414.

Cawsey, T.F., and Wedley, W.C. (1979). Labor turnover costs: Measurement and control. *Personnel Journal*, **2**, 90-95.

Cleary, T.A., and Hilton, T.I. (1968). Test bias: Predictor of grades of Negro and white students in integrated colleges. *Journal of Educational Measurement*, **5**, 115-124.

Cole, N.S. (1973). Bias in selection. *Journal of Educational Measurement*, **10**, 237-255.

Corts, D.B., Muldrow, T.W., and Outerbridge, A.N. (1977). *Research base for the written portion of the Professional and Administrative Career Examination (PACE): Prediction of job success for customs inspectors* (TS-77-4). (NTIS # PB 280620).

Cronbach, L.J. (1960). *Essentials of psychological testing*. New York: Harper & Row.

Cronbach, L.J., and Gleser, G.C. (1965). *Psychological tests and personnel decisions*. Second Edition. Urbana: University of Illinois Press. Originally published 1957.

Cronbach, L.J., Yalow, E., and Schaeffer, G. (1980). A mathematical structure for analyzing fairness in selection. *Personnel Psychology*, **33**, 693-704.

Cronshaw, S.F., and Alexander, R.A. (1985). One answer to the demand for accountability: Selection utility as an investment decision. *Organizational Behavior and Human Decision Processes*, **35**, 102-118.

Curtis, E.W. (1967). *The application of decision theory and scaling methods to selection test evaluation*. Technical Bulletin STB 67-18. San Diego, CA: U.S. Naval Personnel Research Activity.

Curtis, E.W., and Alf, E.F. (1969). Validity, predictive efficiency, and practical significance of selection tests. *Journal of Applied Psychology*, **53**, 327-337.

Darlington, R.B. (1971). Another look at "cultural fairness." *Journal of Educational Measurement*, **8**, 71-82.

DeAngelo, L.E. (1982). Unrecorded human assets and the "hold up" problem. *Journal of Accounting Research*. **20**, 272-274.

Dittman, D.A., Juris H.A., and Revsine, L. (1976). On the existence of unrecorded human assets: An economic perspective. *Journal of Accounting Research*, **14**, 49-65.

Dittman, D.A., Juris, H.A., and Revsine, L. (1980). Unrecorded human assets: A survey of accounting firms' training programs. *The Accounting Review*, **55**, 640-648.

Doppelt, J.E., and Bennett, G.K. (1953). Reducing the cost of training satisfactory workers by using tests. *Personnel Psychology*, **6**, 1-8.

Dreher, G.F., and Sackett, P.R. (1983). *Perspectives on employee staffing and selection*. Homewood, IL: Irwin.

Driessnack, C.H. (1979). Financial impact of effective human resources management. *Personnel Administrator*, **23**, 62-66.

Dubois, B. (1966). A test-dominated society: China 1115 B.C.-1905 A.D. In A. Anastasi (Ed.), *Testing problems in perspective*. Washington, DC: American Council on Education.

Dunnette, M.D., and Borman, W.C. (1979). Personnel selection and classification system. *Annual Review of Psychology*, **30**, 477-525.

Dyl, E.A., and Keaveny, T.J. (1983). Cost minimization in staffing. *Human Resource Planning*, **6**, 103-113.

Eaton, N.K., Hanser, L.M., and Shields, J.L. (1986). Validating selection tests against job performance. In J. Zeidner (Ed.), *Human productivity enhancement: Organizations, personnel and decision making*. (pp. 382-438). New York: Praeger.

Eaton, N.K., Wing, H., and Mitchell, K.J. (1985). Alternate methods of estimating the dollar value of performance. *Personnel Psychology*, **38**, 27-40.

Edwards, W. (1966). Behavioral decision theory. *Annual Review of Psychology*, **18**, 473-498.

Einhorn, H.J., and Bass, A.R. (1971). Methodological considerations relevant to discrimination in employment testing. *Psychological Bulletin*, **75**, 261-269.

Equal Employment Opportunity Commission. Civil Service Commission, Department of Labor, and Department of Justice (1978). Adoption by four agencies of Uniform Guidelines on Employee Selection Procedures. *Federal Register*, **43**, 38290-38315.

Evans, D.W. (1940). Individual productivity differences. *Monthly Labor Review*, **50**, 338-341.

Fernandez, R., and Garfinkle, J. (1985). *Setting enlistment standards and matching recruits to jobs using job performance criteria*. R-3067-MIL. Santa Monica, CA: The Rand Corporation.

Flamholtz, E.G. (1971). A model for human resource valuation: A stochastic process with service awards. *Accounting Review*, **46** 253-267.

Flamholtz, E.G. (1974). *Human resource accounting*. Encino, CA: Dickenson.

Friedman, T., and Williams, E.B. (1982). Current use of tests for employment. In A.K. Wigdor and W.R. Garner (Eds.), *Ability testing: Uses, consequences, and controversies* (Part II, pp. 99-169). Washington, DC: National Academy Press.

Gaudet, F.J. (1960). *Labor turnover: Calculation and cost*. AMA Research Study No. 39. New York: American Management Association.

Ghiselli, E.E. (1956). Dimensional problems of criteria. *Journal of Applied Psychology*, **40**, 1-4.

Ghiselli, E.E. (1959). The generalization of validity. *Personnel Psychology*, **12**, 397-402.

Ghiselli, E.E. (1966). *The validity of occupational aptitude tests*. New York: Wiley.

Ghiselli, E.E. (1973). The validity of aptitude tests in personnel selection. *Personnel Psychology*, **26**, 461-477.

Greer, O.L. (1986). Estimation of the standard deviation of job performance. A comparative study of two behaviorally based methods with a cost-accounting-based approach. Unpublished doctoral dissertation, University of Colorado, Boulder.

Greer, O.L., and Cascio, W.F. (1987). Is cost accounting the answer? Comparison of two behaviorally based methods for estimating the standard deviation of job performance in dollars with a cost-accounting-based approach. *Journal of Applied Psychology*, **72**, 588-595.

Gross, A.L., and Su, W. (1975). Defining a "fair" or "unbiased" selection model: A question of utility. *Journal of Applied Psychology*, **60**, 345-351.

Haney, W. (1981). Validity, vaudeville and values. A short history of social concerns over standardized testing. *American Psychologist*, **36**, 1021-1034.

Harris, R.N. (1967, March). *A model sampling experiment to evaluate two methods of test selection*. (Research Memorandum, 67-2), Statistical Research and Analysis Division. Washington, DC: U.S. Army Behavior and Systems Research Laboratory.

Hekimian, J.S., and Jones, C.H. (1967, January-February). Put people on a balance sheet. *Harvard Business Review*, **45**, 107-113.

Heyne, P. (1988). *Microeconomics*. Chicago: Science Research Associates.

Hirsh, H.R., Schmidt, F.L., and Hunter, J.E. (1986). Estimation of employment validities by less experienced judges. *Personnel Psychology*, **39**, 337-344.

Horst, P. (1954). A technique for the development of a differential prediction battery. *Psychological Monographs*, **68** (9, Whole No. 380).

Horst, P. (1956a). Multiple classification by the method of least squares. *Journal of Clincal Psychology*, **12**, 3-16.

Horst, P. (1956b). Optimal test length for maximum differential predictions. *Psychometrica*, **21**, 51-66.

Howard, R.A. (1966). *Proceedings of the Fourth International Conference on Operational Research*. New York: Wiley.

Howard, R.A., Matheson, J.E., and North, D.W. (1972). The decision to seed hurricanes. *Science*, **176**, 1191-1202.

Hull, C.L. (1928). *Aptitude testing*. Yonkers, NY: World Book Co.

Hunter, J.E. (1983). *Validity generalization for 12,000 jobs: An application of job classification and validity generalization analysis to the General Aptitude Test Battery (GATB).* USES Test Research Report No. 45. Washington, DC: U.S. Employment Service, U.S. Department of Labor.

Hunter, J.E. (1984). *The validity of the Armed Forces Vocational Aptitude Battery (ASVAB) High School Composite.* Report for Research Applications, Inc., in partial fulfillment of DoD Contract No. F41689-83-C-0025.

Hunter, J.E. (1985). *Differential validity across jobs in the military.* Report for Research Applications, Inc., in partial fulfillment of DoD Contract No. F41689-83-C-0025.

Hunter, J.E., Crosson, J.J., and Friedman, D.H. (1985). *The validity of the Armed Services Vocational Aptitude Battery for civilian and military job performance.* Rockville, MD: Research Applications, Inc.

Hunter, J.E., and Hunter, R.F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin*, **96**, 72-98.

Hunter, J.E., and Schmidt, F.L. (1982). Fitting people to jobs: The impact of personnel selection on national productivity. In E.A. Fleishman and M.D. Dunnette (Eds.), *Human performance and productivity: Vol. 1: Human capability assessment.* Hillsdale, NJ: Erlbaum.

Hunter J.E., and Schmidt, F.L. (1983). Quantifying the effects of psychological interventions on employee job performance and workforce productivity. *American Psychologist*, **38**, 473-478.

Hunter, J.E., Schmidt, F.L., and Coggin, T.D. (1988). Problems and pitfalls in using capital budgeting and financial accounting, techniques in assessing the utility of personnel programs. *Journal of Applied Psychology*, **73**, 522-528.

Hunter, J.E., Schmidt, F.L., and Judiesch, M.K. (April 1989). *Individual differences in output variability as a function of job complexity.* In press.

Janz, J.T., and Dunnette, M.D. (1977). An approach to selection decisions: Dollars and sense. In J.R. Hackman et al. (Eds.), *Perspectives on performance in organizations* (pp. 119-126). New York: McGraw-Hill.

Johnson, C.D., and Zeidner, J. (1989, September). *Classification utility: Measuring and improving benefits in matching personnel to jobs.* IDA Paper P-2240. Alexandria, VA: Institute for Defense Analyses.

Katzell, R.B., and Gazzo, R.A. (1983). Psychological approaches to productivity improvement. *American Psychologist*, **38**, 468-472.

Kelley, T.L. (1923). *Statistical methods.* New York: MacMillan.

King, L.M., Hunter, J.E., and Schmidt, F.L. (1980). Halo in a multidimensional forced choice evaluation scale. *Journal of Applied Psychology*, **65**, 507-516.

Klemmer, E.T., and Lockhead, G.R. (1962). Productivity and errors in two keying tasks: A field study. *Journal of Applied Psychology*, **46**, 401-408.

Landy, F.J., Farr, J.L., and Jacobs, R.R. (1982). Utility concepts in performance measurement. *Organizational Behavior and Human Performance*, **30**, 15-40.

Lawshe, C.H. (1948). *Principles of personnel tests*. New York: McGraw-Hill.

Lawshe, C.H. (1952). What can industrial psychology do for small business? *Personnel Psychology*, **5**, 31-34.

Lee, R., and Booth, J.M. (1974). A utility analysis of a weighted application blank designed to predict turnover for clerical employees. *Journal of Applied Psychology*, **59**, 516-518.

Lev, B., and Schwartz, A. (1971). On the use of the economic concept of human capital in financial statements. *Accounting Review*, **46**, 103-112.

Maier, M.H., and Fuchs, E.F. (1972, September). Development and evaluation of a new ACB and aptitude area system. Technical Research Note 239. Alexandria, VA: U.S. Army Research Institute.

Mason, R.O., and Flamholtz, E.C. (1978). Human resources management. In J.J. Moder and S.E. Elmaghraby (eds.), *Handbook of Operations Research, Vol. 2*, 92-110. New York: Van Nostrand Reinhold.

Matheson, J.E. (1969). Decision analysis practice: Examples and insights. In *OR 69: Proceedings of the Fifth International Conference on Operational Research*. Venice: Tavistock Publications.

Mathieu, J.E., and Le [illegible], R.L. (1987). Applying utility concepts to a training program in supervisory skills: A time-based approach. *Academy of Management Journal*, **30**, 316-335.

McCormick, E.J., and Tiffin, J. (1974). *Industrial psychology*. Englewood Cliffs, NJ: Prentice-Hall.

McDaniel, M.A., Schmidt, F.L., Raju, N.S., and Hunter, J.E. (1986). Interpreting the results of meta-analytic research: a comment on Schmitt, Gooding, Abe, and Kirsch (1984). *Personnel Psychology*, **39**, 141-148.

McHenry, J.J. (1987, April). *Project A validity results: The relationship between predictor and criterion domains*. Paper presented at the Society for Industrial and Organizational Psychology Annual Conference, Atlanta, GA.

McKillip, R.H., Trattner, M.H., Corts, D.B., and Wing, H, (1977). *The professional and administrative career examination: Research and development* (PRR-77-1). Washington, DC: U.S. Civil Service Commission, Personnel Research and Development Center. (NTIS # PB 268780/AS).

Mirvis, P.H., and Macy, B.A. (1976). Human resource accounting: A measurement perspective. *Academy of Management Review*, **1**, 74-83.

Murphy, K.R. (1986). When your top choice turns you down: Effect of rejected offers on the utility of selection tests. *Psychological Bulletin*, **99**, 133-138.

Naylor, J.C., and Shine, L.C. (1965). A table for determining the increase in mean criterion score obtained by using a selection device. *Journal of Industrial Psychology*, **3**, 33-42

Nie, N.H., Hull, H.C., Jenkins, J.G., Steinbrenner, K., and Brent, D.H. (1975). *Statistical package for the social sciences*. New York: McGraw-Hill.

O'Connor, E.J., Wexley, K.N., and Alexander, R.A. (1975). Single-group validity: Fact or fallacy? *Journal of Applied Psychology*, **60**, 352-355.

O'Leary, B.S., and Trattner, M.H. (1977). *Research base for the written test portion of the Professional and Administrative Career Examination (PACE): Prediction of job performance for internal revenue office* (TS-77-6). Washington, DC: U.S. Civil Service Commission, Personnel Research and Development Center. (NTIS #PB 274579/AS).

Paperman, J.B. (1977). The current status of human resources accounting. *The Woman CPA*, January 1977, 21-23.

Pearlman, K., Schmidt, F.L, and Hunter, J.E. (1980). Validity generalization results for tests used to predict proficiency and training criteria in clerical occupations. *Journal of Applied Psychology*, **65**, 373-406.

Petersen, N.S., and Novick, M.R. (1976). An evaluation of some models for culture-fair selection. *Journal of Educational Measurement*, **13**, 3-29.

Pyle, W.C. (1970, September). Human resource accounting. *Financial Analysts Journal*, **10**, 68-78.

Raiffa, H. (1968). *Decision analysis: Introductory lectures on choices under uncertainty*. Reading, MA: Addison-Wesley.

Reilly, R.R., and Smither, J.W. (1985). An examination of two alternative techniques to estimate the standard deviation of job performance in dollars. *Journal of Applied Psychology*, **70**, 651-661.

Rich, J.R., and Boudreau, J.W. (1987). The effects of variability and risk in selection utility analysis: An empirical comparison. *Personnel Psychology*, **40**, 55-84.

Roche, W.J., Jr. (1961). The Cronbach-Gleser utility function in fixed-treatment employee selection. *Dissertation Abstracts International*, **22**, 4413. (University Microfilms No. 62-1570). Portions reproduced in L.J. Cronbach and G.C. Gleser (Eds.), *Psychological tests and personnel decisions* (2nd ed). Urbana: University of Illinois Press, 1965, pp. 254-266.

Rothe, H.F. (1946). Output rates among butter wrappers: II. Frequency distributions and a hypothesis regarding the "restriction of output." *Journal of Applied Psychology*, **30**, 320-327.

Rothe, H.F. (1947). Output rates among machine operators: I. Distributions and their reliability. *Journal of Applied Psychology*, **31**, 484-489.

Rothe, H.F. (1951). Output rates among chocolate dippers. *Journal of Applied Psychology*, **25**, 94-97.

Rothe, H.F. (1970). Output rates among welders: Productivity and consistency following removal of a financial incentive system. *Journal of Applied Psychology*, **54**, 549-551.

Rothe, H.F. (1978). Output rates among industrial employees, *Journal of Applied Psychology*, **63**, 40-46.

Rothe, H.F., and Nye, C.T. (1958). Output rates among coil winders. *Journal of Applied Psychology*, **42**, 182-186.

Rothe, H.F., and Nye, C.T. (1959). Output rates among machine operators: II. Consistency related to methods of pay. *Journal of Applied Psychology*, **43**, 417-420.

Rothe, H.F., and Nye, C.T. (1961). Output rates among machine operators: III. A nonincentive situation in two levels of business activity. *Journal of Applied Psychology*, **45**, 50-54.

Sadacca, R., deVerra, M.V., DiFazio, A.S., and White, L.A. (1986, August). *Weighting performance constructs in composite measures of job performance.* Paper presented at the American Psychological Association, Annual Meeting, Washington, DC.

Sands, W.A. (1973a). A method for evaluating alternative recruiting--selection strategies. The CAPER model. *Journal of Applied Psychology*, **57**, 222-227.

Sands, W.A. (April 1973b). *A bivariate normal version of the cost of attaining personnel requirements model.* WTR 73-18. Washington, DC: Naval Personnel Research and Development Laboratory.

Savich, R.S., and Ehrenreich, K.P. (1976). Cost/benefit analysis of human resource accounting alternatives. *Human Resource Management*, **15**, 7-18.

Sawyer, R.L., Cole, N.S., and Cole, J.W.L. (1976). Utilities and the issue of fairness in a decision-theoretic model for selection. *Journal of Educational Measurement*, **13**, 59-76.

Schmidt, F.L., Gast-Rosenberg, J., and Hunter, J.E. (1980). Validity generalization results for computer programmers. *Journal of Applied Psychology*, **65**, 643-661.

Schmidt, F.L., and Hoffman, B. (1973). Empirical comparison of three methods of assessing utility of a selection device. *Journal of Industrial and Organizational Psychology*, **1**, 13-22.

Schmidt, F.L., and Hunter, J.E. (1981). Employment testing: Old theories and new research findings. *American Psychologist*, **36**, 1128-1137.

Schmidt, F.L., and Hunter, J.E. (1983). Individual differences in productivity: An empirical test of estimates derived from studies of selection procedures utility. *Journal of Applied Psychology*, **68**, 407-414.

Schmidt, F.L., Hunter, J.E., and Dunn, W.L. (1987, November). *Potential utility increases from adding new tests to the Armed Services Vocational Aptitude Battery (ASVAB).* Contract No. Delivery Order 0053. San Diego, CA: U.S. Navy Personnel Research and Development Center.

Schmidt, F.L., Hunter, J.E., Croll, P.R., and McKenzie, R.C. (1983). Estimation of employment test validities by expert judgment. *Journal of Applied Psychology*, **68**, 590-601.

Schmidt, F.L., Hunter, J.E., and Larson, M. (1988, August). *General cognitive ability vs. general and specific aptitudes in the prediction of training performance: Some preliminary findings.* Contract No. Delivery Order 0053. San Diego, CA: U.S. Navy Personnel Research and Development Center.

Schmidt, F.L., Hunter, J.E., McKenzie, R.C., and Muldrow, T.W. (1979). The impact of valid selection procedures on workforce productivity. *Journal of Applied Psychology*, **64**, 609-626.

Schmidt, F.L., Hunter, J.E., Outerbridge, A.N., and Trattner, M.H. (1986). The economic impact of job selection methods on size, productivity, and payroll costs of the federal work force: An empirically based demonstration. *Personnel Psychology*, **39**, 1-29.

Schmidt, F.L., Hunter, J., and Pearlman, K. (1982). Assessing the economic impact of personnel programs on work-force productivity. *Personnel Psychology*, **35**, 333-347.

Schmidt, F.L., Hunter, J.E., Pearlman, K., and Shane, G.S. (1979). Further tests of the Schmidt-Hunter validity generalization model. *Personnel Psychology*, **32**, 257-281.

Schmidt, F.L., Mack, M.J., and Hunter, J.E. (1984). Selection utility in the occupation of U.S. park ranger for three modes of test use. *Journal of Applied Psychology*, **69**, 490-497.

Schmitt, N., Gooding, R.Z., Noe, R.D., and Kirsch, M. (1984). Meta-analysis of validity studies published between 1964 and 1982 and the investigation of study characteristics. *Personnel Psychology*, **37**, 407-422.

Simonet, J.K. (1984). *The convergent validity of methods of estimating the standard deviation of job performance in dollars.* Dissertation, University of Georgia, published in 1986, Ann Arbor, MI: University Microfilms International.

Society for Industrial and Organizational Psychology (1987). *Principles for the validation and use of personnel selection procedures.* College Park, MD: Society for Industrial and Organizational Psychology.

Stead, W.H., and Shartle, C.L. (1940). *Occupational counseling techniques.* New York: American Book.

Taylor, H.C., and Russell, J.T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: Discussion and tables. *Journal of Applied Psychology*, **23**, 565-578.

Thorndike, R.L. (1971). Concepts of culture-fairness. *Journal of Educational Measurement*, **8**, 63-70.

Tiffin, J. (1947). *Industrial psychology* (2nd ed.). New York: Prentice-Hall.

Trattner, M.H., Corts, D.B., van Rijn, P.P. and Outerbridge, A.N. (1977). *Research base for the written test portion of the Professional and Administrative Career Examination (PACE): Prediction of job performance for claims authorizers in the Social Insurance Claims Examining occupation* (TS-77-3). Washington, DC: U.S. Civil Service Commission, Personnel Research and Development Center. (NTIS #PB 273118/AS).

Tsay, J.J. (1977). Human resource accounting. A need for relevance. *Financial Analysis Journal*, **58**, 33-36.

Van Naersson, R.F. (1963). Selectie van chauffeurs, Gronigen: Wolters. Portions translated in L.J. Cronbach and G.C. Gleser (eds.). *Psychological tests and personnel decisions*. Urbana: University of Illinois Press, 1965, pp. 273-290.

Viteles, M.S., (1932). *Industrial psychology*. New York: Norton.

Von Neumann, J., and Morgenstern, O. (1947). *Theory of games and economic behavior* (2nd ed.). Princeton, NJ: Princeton University Press.

Wald, A. (1950). *Statistical decision functions*. New York: Wiley.

Ward, D.L. (1982). The $34,000 lay-off. *Human Resource Planning*, **8**, 35-41.

Wechsler, D. (1952). *Range of human capacities* (2nd ed.). Baltimore, MD: Williams and Wilkins.

Weekley, J.A., Frank, B., O'Connor, E.J., and Peters, L.H. (1985). A comparison of three methods of estimating the standard deviation of performance in dollars. *Journal of Applied Psychology*, **70**, 122-126.

Wigdor, A.K., and Garner, W.E. (Eds.). (1982). *Ability testing: Uses, consequences and controversies*. Part I. Report of the Committee. Washington, DC: National Academy Press.

Wing, H. (1977). *Status of Test Usage in FY 77*. Personnel Research and Development Center, Test Services Section, Technical Note 77-2. Washington, DC: U.S. Civil Service Commission.

Zedek, S., and Cascio, W.F. (1984). Psychological issues in personnel decisions. *Annual Review of Psychology*, **35**, 461-518.

Zeidner, J. (1987, April). *The validity of selection and classification procedures for predicting job performance.* IDA Paper P-1977. Alexandria, VA: Institute for Defense Analyses.

Zeidner, J., and Johnson, C.D. (1989, September). The economic benefits of predicting job performance. IDA Paper P-2241. Alexandria, VA: Institute for Defense Analyses.

A217 608

AD NUMBER	**********	E501187 **********
FIELD 2:	FLD/GRP(S)	150100
FIELD 3:	ENTRY CLASS.	U
FIELD 4:	NTIS PRICE	HC MF
FIELD 5:	SOURCE NAME	[INSTITUTE FOR DEFENSE ANALYSES ALEXANDRIA VA
FIELD 6:	UNCLASS. TITLE	[T]HE [U]TILITY OF [S]ELECTION FOR [M]ILITARY AND [C]IVILIAN [J]OBS.
FIELD 7:	CLASS. TITLE	
FIELD 8:	TITLE CLASS.	U
FIELD 9:	DESCRIPTIVE NOTE	[F]INAL REPT. [J]UN 88-[J]UL 89,
FIELD 10:	PERSONAL AUTHORS	[Z]EIDNER, [J]OSEPH; [J]OHNSON, [C]ECIL [D].
FIELD 11:	REPORT DATE	JUL 89
FIELD 12:	PAGINATION	233P
FIELD 13:	PROCESSING LEVEL	
FIELD 14:	REPORT NUMBER	[IDA]-[P]-2239
FIELD 15:	CONTRACT NUMBER	[MDA]903-84-[C]-0031
FIELD 16:	PROJECT NUMBER	
FIELD 17:	TASK NUMBER	
FIELD 18:	MONITOR ACRONYM	[IDA/HQ], [SBI
FIELD 19:	MONITOR SERIES	89-34657, [AD]-[E]501 187
FIELD 20:	REPORT CLASS	U
FIELD 21:	SUPPLEMENTARY NOTE	
FIELD 22:	ALPHA LIMITATIONS	[A]VAILABILITY CONTROLLED BY [IDA], [A]TTN: [TIS], [A]LEXANDRIA, [VA]22311. [A]NNOUNCEMENT ONLY; DOCUMENT WILL BE MADE AVAILABLE FROM [DTIC]AFTER PROCESSING.
FIELD 23:	DESCRIPTORS	[MILITARY TRAINING], [PERFORMANCE(ENGINEERING)], [JOB ANALYSIS], [COST EFFECTIVENESS], [MANPOWER], [PERFORMANCE TESTS], [MILITARY PERSONNEL], [APTITUDE TESTS].
FIELD 24:	DESCRIPTOR CLASS.	U
FIELD 25:	IDENTIFIERS	[LPN]-[IDA]-[T]-[D]2-435, [SBI]1, [F]ISCAL YEAR 1990, [S]ELECTION UTILITY.
FIELD 26:	IDENTIFIER CLASS.	U
FIELD 27:	ABSTRACT	[T]HE MAJOR PURPOSE OF THIS REPORT IS TO PROVIDE MILITARY POLICYMAKERS WITH PROCEDURES FOR DEVELOPING AND EVALUATING REALISTIC ESTIMATES OF COSTS AND BENEFITS OF ALTERNATIVE MANPOWER SELECTION AND CLASSIFICATION POLICIES. [S]UCH ESTIMATES ARE NEEDED TO MAKE RATIONAL CHOICES IN ALLOCATING SCARCE RESOURCES AMONG STRATEGIES FOR IMPROVING ORGANIZATIONAL PRODUCTIVITY. [T]HIS REPORT TRACES THE TECHNICAL DEVELOPMENT OF CURRENT DECISION THEORETIC SELECTION UTILITY MODELS. [T]HE DESCRIPTION OF SELECTION IS EXTENDED TO INTRODUCE MORE COMPLEX CLASSIFICATION DECISION SITUATIONS THAT MATCH INDIVIDUALS AND JOBS TO MAXIMIZE AGGREGATE PERFORMANCE. [A]N OVERVIEW OF THE CURRENT MILITARY SYSTEM FOR SELECTING AND CLASSIFYING MANPOWER IS PRESENTED ALONG WITH A DISCUSSION OF HOW EXCLUSIVE FOCUS ON PREDICTING VALIDITY REDUCES THE EFFICIENCY OF THE [ASVAB]AS A CLASSIFICATION TOOL.
FIELD 28:	ABSTRACT CLASS.	U
FIELD 29:	INITIAL INVENTORY	2
FIELD 30:	ANNOTATION	
FIELD 31:	SPECIAL INDICATOR	
FIELD 32:	REGRADE CATEGORY	
FIELD 33:	LIMITATION CODES	1 2
FIELD 34:	SOURCE SERIAL	[F
FIELD 35:	SOURCE CODE	179350
FIELD 36:	DOCUMENT LOCATION	
FIELD 37:	CLASSIFIED BY	
FIELD 38:	DECLASSIFY ON	
FIELD 39:	DOWNGRDE TO CONF ON	
FIELD 40:	GEOPOLITICAL CODE	5108
FIELD 41:	TYPE CODE	W
FIELD 42:	IAC ACCESSION NO.	
FIELD 43:	IAC DOCUMENT TYPE	
FIELD 44:	IAC SUBJECT TERM	

FIELD 45: EXTENDED BY
FIELD 46: REVIEW ON DATE
FIELD 47: REASON CODE
FIELD 48: SBIE SITE SYMBOLS IDAHO34657
AD NUMBER ********** E501187 **********